Complete sequence of the human mucin MUC4: a putative cell membrane-associated mucin

Nicolas MONIAUX*, Séverine NOLLET*, Nicole PORCHET*†, Pierre DEGAND*†, Anne LAINE* and Jean-Pierre AUBERT*†[1]

*Unité 377 INSERM, Place de Verdun, 59045 Lille Cedex, France, and †Laboratoire de Biochimie et de Biologie Moléculaire de l'Hôpital C. Huriez, 59037 Lille Cedex, France

The *MUC4* gene, which encodes a human epithelial mucin, is expressed in various epithelial tissues, just as well in adult as in poorly differentiated cells in the embryo and fetus. Its N-terminus and central sequences have previously been reported as comprising a 27-residue peptide signal, followed by a large domain varying in length from 3285 to 7285 amino acid residues. The present study establishes the whole coding sequence of MUC4 in which the C-terminus is 1156 amino acid residues long and shares a high degree of similarity with the rat sialomucin complex (SMC). SMC is a heterodimeric glycoprotein complex composed of mucin (ascites sialoglycoprotein 1, ASGP-1) and transmembrane (ASGP-2) subunits. The same organization is found in MUC4, where the presence of a GlyAspProHis proteolytic site may cleave the large precursor into two subunits, MUC4α and MUC4β. Like ASGP-2, which binds the receptor tyrosine kinase p185^{neu}, MUC4β possesses two epidermal growth factor-like domains, a transmembrane sequence and a potential phosphorylated site. MUC4, the human homologue of rat SMC, may be a heterodimeric bifunctional cell-surface glycoprotein of 2.12 μm. These results confer a new biological role for MUC4 as a ligand for ErbB2 in cell signalling.

Key words: ascites sialoglycoprotein, epidermal growth factor-like domain, epithelial tissue, membrane glycoprotein.

INTRODUCTION

Originally, mucins were defined as the major highly glycosylated glycoproteins that composed the slimy and viscous secretion that covers epithelial surfaces, the mucus. These mucins, now called epithelial mucins, are thought to play an important role in the protection of the epithelial cells, and they have been implicated in epithelial renewal and differentiation [1,2]. Currently, nine human mucin genes have been identified, designated *MUC1–4*, *MUC5B*, *MUC5AC* and *MUC6–8* [3–11]. Four of these *MUC* genes, *MUC6*, *MUC2*, *MUC5AC* and *MUC5B*, are clustered between *HRAS* and *IGF2* on chromosome 11 in p15.5 [12]. These genes, which exhibit some sequence similarities with the cysteine-rich domains of the pro-von Willebrand factor, have been proposed to be derived from a common ancestral gene [13], and are believed to be the gel-forming mucins. MUC7, which does not show any cysteine-rich domain, is a soluble secreted mucin. The human epithelial mucin MUC3 is expressed in the small intestine, in both goblet cells and villus columnar cells [14,15]. MUC3 exhibits one epidermal growth factor (EGF)-like domain of which a precise role is still unknown [16]. Until now, no transmembrane region has been identified in human MUC3, although its mouse and rat homologues have one [17,18]. MUC1, which is expressed on the apical surface of most secretory epithelia [19], was the first human epithelial membrane-bound mucin identified.

The first partial cDNA from *MUC4* was isolated in our laboratory from a human tracheobronchial cDNA library [6]. MUC4 is expressed in a variety of tissues, including trachea and the bronchial area, cervix, stomach, small intestine and colon [15–20]. Like MUC3, MUC4 is not restricted to goblet cells. It is also expressed in the ciliated cells of trachea and bronchi and in absorptive cells of intestinal mucosa. Recently, the genomic organization of the 5′ region and the central part of the *MUC4* gene was determined [21]. The first exon codes the signal peptide of 27 residues and shares a high degree of similarity with that of ascites sialoglycoprotein 1 (ASGP-1), part of the rat sialomucin complex (SMC). The second exon is a large one and contains a unique sequence (951 residues) that is followed by a long tandem-repeat (TR) domain. This TR domain varies in length from 2334 to 6334 amino acid residues. This variation is due to variable number of tandem repeats (VNTR) polymorphism. The rat SMC is a well-characterized membrane-associated mucin [22]. SMC was originally isolated and characterized as a heterodimeric glycoprotein complex from highly metastatic 13762 rat mammary adenocarcinoma ascites cells, in which the mucin subunit ASGP-1 is the major detectable glycoprotein [23]. The other subunit of SMC, ASGP-2, which contains two EGF-like domains [24], has been shown to act as a ligand for the tyrosine kinase p185^{neu} [25]. The present study establishes the whole deduced coding sequence of the MUC4 C-terminus, which is a 1156-residue peptide. The MUC4 C-terminus, which shares a high degree of similarity with SMC, also possesses two EGF-like domains, a potential transmembrane sequence, a putative GlyAspProHis (GDPH) proteolytic cleavage site, two domains rich in potential N-glycosylation sites and two cysteine-rich domains. Our results allow us to conclude that MUC4 is the human homologue of rat SMC.

EXPERIMENTAL

Library screening

Total RNA was extracted from a human colon mucosa using the guanidinium isothiocyanate/CsCl method [26] and used as a template for cDNA synthesis. All details concerning double-stranded cDNA synthesis and cloning into λgt11 vector were as described by the commercial supplier, Amersham (Saclay,

Abbreviations used: SMC, sialomucin complex; EGF, epidermal growth factor; ASGP, ascites sialoglycoprotein; TR, tandem repeat; RACE, rapid amplification of cDNA ends; RT-PCR, reverse-transcriptase PCR.

[1] To whom correspondence should be addressed (e-mail jpa@lille.inserm.fr).

The nucleotide sequence data reported is in the EMBL Nucleotide Sequence Database under the accession number AJ010901.

France). Nitrocellulose membranes (Schleicher and Schüll, Ceralabo, Ecquevilly, France) were used to obtain plaque lifts. These membranes were prehybridized and hybridization was performed with 2.5×10^5 c.p.m./membrane at 42 °C overnight. Inserts of positive phages were subcloned into pBluescript KS$^+$ vector.

Cloning in pBluescript KS$^+$

Restriction enzyme digestions (*Bam*HI, *Acc*I, *Pst*I) were performed under standard conditions with the appropriate buffer on the cosmid genomic clone, LEA2, isolated previously [21]. The different fragments obtained were subcloned into pBluescript KS$^+$ vector from Stratagene (Ozyme, Saint Quentin en Yvelines, France). Subclones were sequenced using the T3 and T7 vector primers, and sequences were analysed with the GenBank™ database.

Plasmid DNA purification

Qiaprep Spin Plasmid Kit (Qiagen, Courtaboeuf, France) was used according to the manufacturer's instructions.

3′-Rapid amplification of cDNA ends (RACE)-PCR procedure

Total RNA from human colon mucosa was extracted as described previously [26]. Advantage™ RT-for-PCR kit (Clontech, Heidelberg, Germany) was used to synthesize first-strand cDNA from 1 μg of RNA using the oligo (dT)-anchor primer of the 5′/3′-RACE kit (Boehringer Mannheim, Roche Diagnostics, Meylan, France). Expand long PCR was performed using Expand™ Long Template PCR System (Boehringer Mannheim) with the sense primer NAU 491 (5′-AGCAGGCCGAGTCTTGGATTA-3′), and as antisense primer the PCR anchor primer of the 5′/3′-RACE kit was used. The PCR amplification reaction mixture (50 μl) contained 5 μl of cDNA, 10 mM sodium dNTPs, 0.4 μM of each primer, 5 μl of 10 × Expand™ Long Template PCR buffer 3, 0.75 mM $MgCl_2$ and 2.5 units of enzyme mixture. The PCR was performed using a Perkin-Elmer Thermal Cycler Gene Amp® PCR System 9700. PCR parameters were 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, annealing at 60 °C for 45 s and elongation at 71 °C for 4 min, of which the 20 last cycles had their elongation time extended by 40 s for each new cycle, followed by a final elongation at 71 °C for 15 min. Nested PCR was carried out using NAU 483 (5′-CTGTTTCTCTACCAGAGCGGT-3′) and the PCR anchor primer. The amplified product was electrophoresed on 1 % TBE (1 × TBE = 45 mM Tris/borate/1 mM EDTA) agarose gel and stained with ethidium bromide. The band was cut out, purified using QIAquick Gel Extraction Kit (Qiagen), and subcloned into the Original TA Cloning® Kit (Invitrogen, Leek, The Netherlands).

Oligonucleotide primers

Oligonucleotide primers used in PCR, RACE-PCR, reverse-transcriptase PCR (RT-PCR) and sequencing experiments were synthesized by MWG-Biotech (Ebersberg, Germany). These primers were: sense NAU 139 (nt 1–22), antisense NAU 363 (nt 425–445), sense NAU 491 (nt 515–535), sense NAU 483 (nt 682–702), antisense NAU 577 (nt 1273–1293), sense NAU 576 (nt 1294–1314), sense NAU 590 (nt 1660–1680), sense NAU 591 (nt 1976–1996), sense NAU 585 (nt 2339–2359), antisense NAU 584 (nt 2582–2602), antisense NAU 555 (nt 2728–2748), sense NAU 586 (nt 2728–2748), antisense NAU 589 (nt 2910–2930), sense NAU 511 (nt 2994–3014), sense NAU 587 (nt 3213–3233), antisense NAU 535 (nt 3302–3322) and antisense NAU 533 (nt 3569–3589).

Sequencing and sequence analyses

Clones were sequenced on both strands by the dideoxy chain-termination method using [α-^{35}S]dATP with Sequenase version 2.0 (Amersham). Sequences were also determined by automatic sequencing, using internal primers with an ABI Prism model 377 XL automatic sequencer and the ABI PRISM dRhodamine terminator cycle sequencing ready reaction kit (Perkin-Elmer, Inc., Courtaboeuf, France) or using the standard vector primers, with a DNA Sequencer model 4000L LI-COR and the SequiTherm Excel™ II long-read Premix DNA Sequencing Kit-LC (TEBU, Le Perray en Yvelines, France).

Analyses of nucleic acid and protein sequence data were performed using PC/GENE Software. The nucleotide sequence reported in this paper has been submitted to the EMBL Databank with accession number AJ010901.

RT-PCR amplification

RNA from human colon mucosa (1 μg) was used to perform single-strand cDNA using the Advantage™ RT-for-PCR kit (Clontech) with a poly(T) primer. PCR was performed with sense NAU 139 and antisense NAU 363 as primer using a Perkin-Elmer Thermal Cycler Gene Amp® PCR System 9700. PCR parameters were 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, annealing at 60 °C for 45 s and extension at 72 °C for 1 min, followed by a final elongation at 72 °C for 15 min. PCR were performed using 2.5 units of *Taq* DNA polymerase (Boehringer Mannheim). The amplified product was electrophoresed on 1 % TBE agarose gel and stained with ethidium bromide.

Northern-blot analysis

RNA from human colon prepared with the improved method for isolation of large RNA was used to perform Northern-blot analysis as described previously [27].

RESULTS

Isolation of the first exons downstream of the 48 bp repetitive sequence

Fragments of the previously isolated genomic clone LEA2 [21] corresponding to the region downstream of the 48 bp TR were subcloned and partially sequenced (Figure 1A). The different sequences obtained were compared with the GenBank™ data base. The 3′ end of the *Acc*I–*Acc*I 2.8 kb fragment called S610 showed 78 % of similarity with a region situated downstream of the rat ASGP-1 tandem repeat. S610 used as a probe and hybridized with a multiple-tissue Northern blot exhibited the same pattern of expression as that obtained with the JER64 probe (results not shown). An antisense oligonucleotide, NAU 363, was chosen from the end of S610 to perform RT-PCR on mRNA extracted from a normal human colon mucosa with the sense oligonucleotide, NAU 139, chosen in the first 21 nucleotides downstream of the 48 bp repetitive sequence. The RT-PCR procedure produced a 446 bp fragment, S1217 (Figure 1B). The sequence determined was analysed and compared with that of the genomic clone LEA2. S1217 showed 100 % identity with the sequence of LEA2, in three exons dispersed along 7 kb of the cosmid clone. The first exon (E3) of 175 bp encodes a domain rich in serine, threonine and proline, the second (E4) is 134 bp long and the third (E5) is at least 137 bp long.

S1217 was used to screen a human colon mucosa cDNA library. One positive clone was isolated and named JER107 (Figure 1C). The JER107 insert consists of 920 bp, of which the

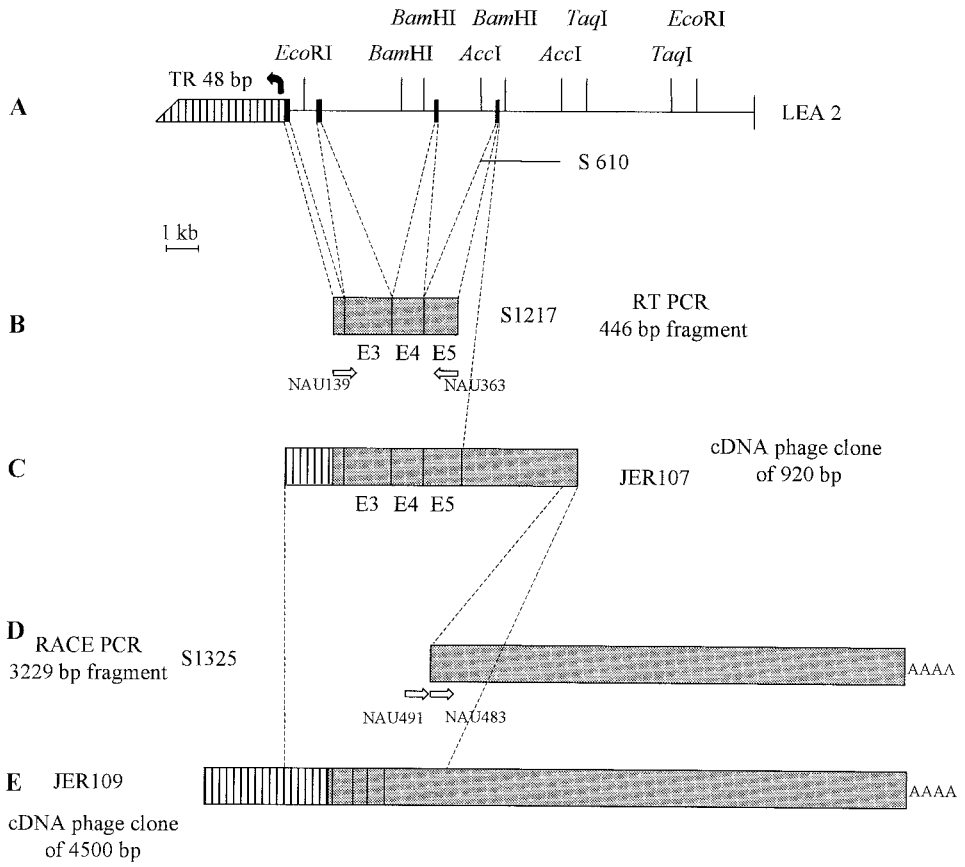

Figure 1 Map of *MUC4* 3′-terminal clones

(**A**) Partial restriction map of LEA2 pWE15 cosmid clone. The fragment called S610 was subcloned into pBluescript KS$^+$ vector. The positions of the three exons, E3, E4 and E5, are indicated by black boxes. Some primers and their directions are indicated (not to scale) by horizontal arrows. (**B**, **C**, **D**, **E**) Different cDNA clones isolated by RT-PCR, library screening and RACE-PCR.

first 70 bp is 48 bp repetitive sequence; the following 446 bp show 100 % identity with the S1217 sequence and extend the exon E5 by 28 bp. Comparison between JER107 and the cosmid LEA2 sequences reveals the presence of at least four introns, I2–I5, of which three contain sequences repeated in tandem. The first is in I3 and consists of the 15 bp TR isolated previously [21]. The second in I4 is a novel 26–32 bp imperfect TR. The third TR in I5 is a 32 bp nearly perfect TR.

Extension of the 3′-terminus cDNA of *MUC4*

One sense primer, NAU 491, was chosen in the 3′ end of the phage clone JER107 to extend the sequence by 3′-RACE-PCR on human colonic mucosa total RNA. One 3396 bp cDNA fragment was obtained. Another sense primer, NAU 483, chosen in JER107, was used to perform a nested PCR on the 3396 bp fragment. One cDNA of 3229 bp was obtained and named S1325 (Figure 1D). The first 169 bp of S1325 show 100 % identity with the 3′ end of JER107. JER107 was also used as a probe to screen the colon cDNA library. One positive clone, called JER109, was isolated (Figure 1E), which overlaps with S1325. It was sequenced in its entirety.

Analysis of the nucleotide and deduced amino acid sequences of the *MUC4* 3′-end cDNA

The compiled nucleotide sequences of the different cDNA clones isolated allowed us to establish the whole coding sequence of the *MUC4* 3′-terminus and its junction with the large 48 bp TR region. The unique sequence downstream of the 48 bp TR consists of a 3468 bp sequence that encodes a 1156-residue peptide (Figure 2) followed by a 3′-untranslated region of 405 bp. This compiled nucleotide sequence shows a high degree of similarity with the C-terminal sequence of the well-characterized rat membrane mucin called SMC, and can be subdivided in 13 regions, which encode 12 distinct domains (Table 1). Structural organizations of the C-termini of both peptides (MUC4 and SMC) are very similar (Figure 3).

The first four domains (CT1–CT4) are separated from the others by a putative GDPH proteolytic cleavage site. An identical proteolytic cleavage site exists in SMC between ASGP-1 and ASGP-2. Except CT1, all the C-terminal MUC4 domains exhibit sequence similarity to the corresponding domains of ASGP-1 or ASGP-2.

CT1 encodes a mucin-like domain comprising 12.5 % serine, 23 % threonine and 16 % proline. This sequence is different from the unique domain rich in serine, threonine and proline of the MUC4 N-terminus described previously [21], or from the 16-amino-acid TR domain.

CT2 encodes a unique non-mucin type sequence, which shows a high degree of similarity with a 3′ region of ASGP-1. The degree of identity with ASGP-1 is higher at the nucleotide level. The similarity between the two molecules is particularly striking, about 60 %, if we consider amino acids 195–284 of MUC4 and amino acids 1972–2061 of ASGP-1 [28].

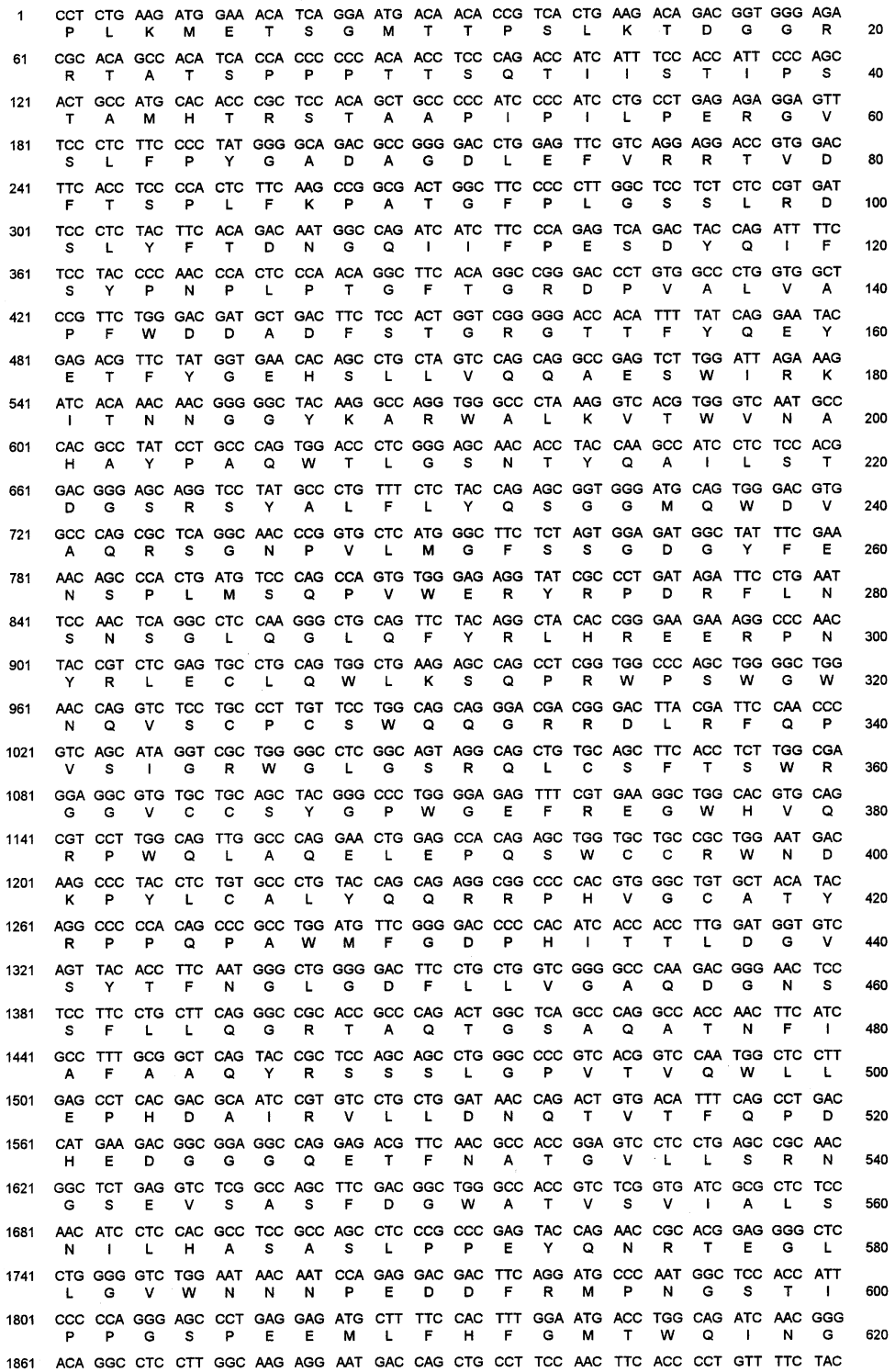

Figure 2 For legend see facing page

CT3 encodes a cysteine-rich domain comprising 11.3% cysteine. The nucleotide sequence of this domain shows 78% similarity with the ASGP-1 sequence, but the two deduced peptides are different. As in CT2, there are several changes in the reading frame. The analysis of this sequence with the GenBank™ data base does not exhibit evidence of similarity with any other cysteine-rich domain and particularly with the cysteine-rich domains found in the 11p15.5 mucin gene family.

CT4 encodes a peptide which shows 64% similarity with ASGP-1 in amino acids 2189–2202. CT5 encodes a large domain that shows 60% similarity with the first subdomain of ASGP-2, which contains 16 putative N-glycosylation sites. CT5 contains

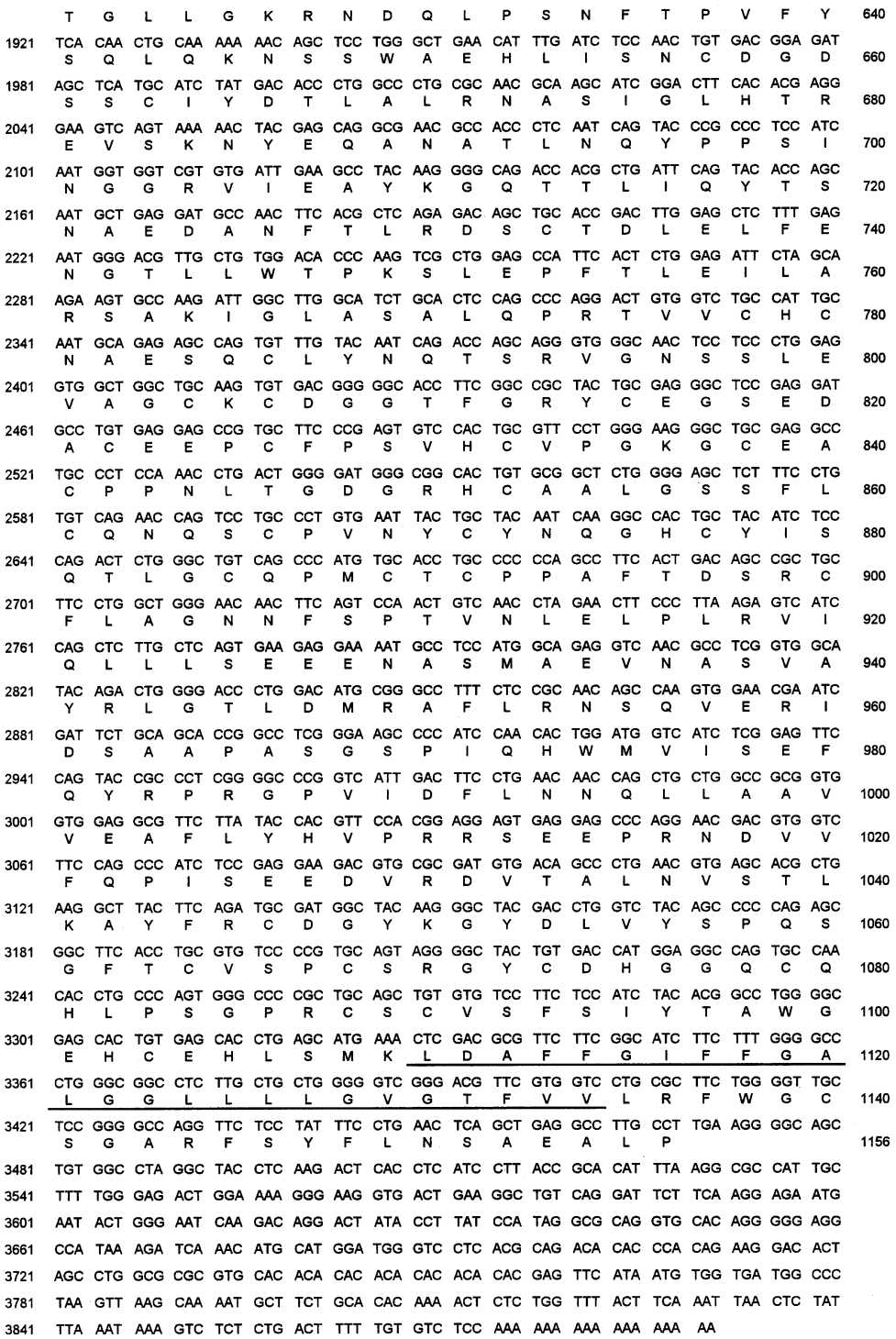

Figure 2 Compiled nucleotide sequences and deduced amino acid sequence of the 3′ terminus of *MUC4*

Nucleotide 1 corresponds to the first nucleotide just downstream of the 48 bp TR sequence. The hydrophobic stretch of amino acid residues is underlined. The nucleotides are numbered on the left and the amino acid residues are numbered on the right side of the Figure.

13 potential N-glycosylated sites, of which 10 are conserved in both peptides.

CT6 encodes another cysteine-rich domain. This domain, which contains 14.5% cysteine residues, shows 65% similarity with a cysteine-rich domain in ASGP-2. In ASGP-2, this region follows the N-glycosylation-rich domain as well. Like the N-glycosylation sites, the cysteines are conserved in both peptides. This domain contains three potential N-glycosylation sites that are also found in ASGP-2. As is the case for the cysteine-rich domain found upstream of the GDPH cleavage site, no similarity

Table 1 Position and characterization of the different domains of the MUC4 C-terminal region and their similarity with the different subunits of rat SMC

Name	Position in nucleotide	Characteristic	Similarity with SMC
CT1	1–168	Mucin-like domain	
CT2	169–912	Unique sequence	ASGP1
CT3	913–1251	Cysteine-rich domain	ASGP1
CT4	1252–1293	Unique sequence	ASGP1
	1288–1299	GDPH cleavage site	GDPH cleavage site
CT5	1294–2331	N-Glycosylated rich domain	ASGP2
CT6	2332–2580	Cysteine-rich domain	ASGP2
CT7	2581–2700	EGF1 domain	ASGP2
CT8	2701–3135	N-Glycosylated rich domain	ASGP2
CT9	3136–3270	EGF2 domain	ASGP2
CT10	3271–3327	Unique sequence	ASGP2
CT11	3328–3401	Transmembrane domain	ASGP2
CT12	3402–3468	Cytoplasmic tail	ASGP2
CT13	3469–3873	3′ Untranslated sequence	

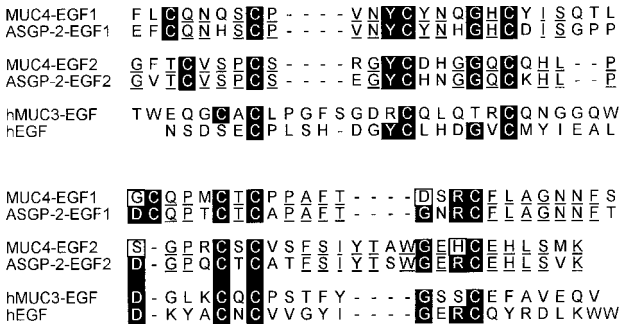

Figure 4 Comparison of EGF-like domains of MUC4β with those of rat ASGP-2, human MUC3 and human EGF

Highly conserved cysteine residues and other essential residues are in white lettering on a dark background and conserved residues between SMC and MUC4 are underlined. Non-conserved essential amino acids are boxed.

exists with the cysteine-rich domains in the *MUC* genes in the 11p15.5 mucin family.

CT7 and CT9 are two EGF-like domains. Comparison between different EGF-like domains is shown in Figure 4. The similarity between both peptides (MUC4 and ASGP-2) is, respectively, 68% for EGF1 and 67% for EGF2. The positions of all the cysteine residues are identical in both molecules, with a $CX_4CX_4CX_5CX_7CX_3CXCX_8C$ motif for EGF1 and a $CX_{17}CX_3CX_4CX_5CX_8CXCX_{12}C$ motif for EGF2 (where X denotes any other residue). Moreover, there is a putative N-glycosylation site in position 863, which is also found in ASGP-2. It is important to note that one aspartic acid and one glycine found in most of the EGF-like domains are replaced, respectively, by one glycine in position 884 and by one aspartic acid in position 887 in MUC4 EGF1. In MUC4 EGF2, one aspartic acid and one arginine are replaced, respectively, by a serine in position 1084 and by a histidine in position 1102. Moreover, MUC4 EGF1 possesses the supplementary cysteine residue found in ASGP-2 EGF1 (second block in Figure 4). The motif found in MUC4 and ASGP-2 EGF-like domains is CX_7CX_3C instead of the $CX_{10}C$ motif that is usually found in the other EGF-like domains. As in ASGP-2, a domain of 147 amino acid residues (CT8) separates the two EGF-like domains. This domain contains four potential N-glycosylation sites that are conserved in ASGP-2.

CT10, encodes a domain that shows 83% similarity with ASGP-2. This domain separates MUC4 EGF2 from a very

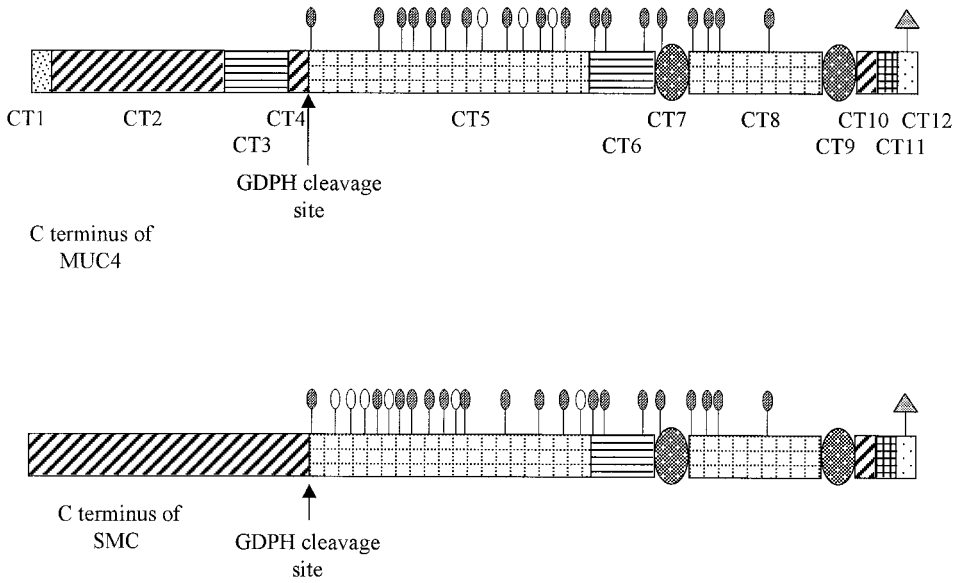

Figure 3 Schematic representation of human MUC4 and rat SMC C-termini

Dense dots, serine/threonine-rich non-repetitive sequence domain; diagonal lines, unique sequence; horizontal lines, cysteine-rich domain; dotted grid, domain rich in potential N-glycosylation sites; hatched ovals, EGF-like domain; solid grid, potential transmembrane sequence; spaced dots, potential cytoplasmic tail; and (on stalks above sequence) hatched ovals, conserved potential N-glycosylation sites; open ovals, non-conserved potential N-glycosylation sites; hatched triangles, potential phosphorylated sites.

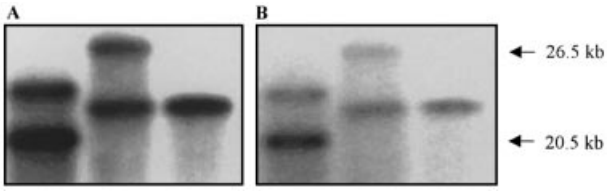

Figure 5 Comparison between Northern-blot patterns obtained with the JER64 and S1325 probes

Total RNA prepared from three individual colons was hybridized with the JER64 (**A**) and S1325 (**B**) probes.

hydrophobic domain (CT11) [29]. CT11 shows 63 % similarity with the transmembrane sequence of ASGP-2.

The last coding region, CT12, shows 55 % similarity with the cytoplasmic tail of ASGP-2. It does not possess the palmitoylation site CXC found in ASGP-2 [24] and MUC1 [30]. This domain contains one tyrosine residue at position 1147. A tyrosine at the same position in the cytoplasmic tail of ASGP-2 is suspected to be a putative phosphorylation site.

CT13, a 3′-untranslated region of 405 bp, contains two potential polyadenylation signals (AAATTAA and AAATAAA). This region does not share any similarity with the 3′-untranslated region of *ASGP-2* except for a 10 CA motif repeated in tandem in both cDNAs.

RNA analysis

The whole 3′-end cDNA fragment (S1325) was used as a probe to hybridize a Northern blot of three individuals' colonic mucosae (prepared with improved method for isolation of large RNA [27]). This fragment revealed the same double bands that were revealed with the JER64 probe (Figure 5).

DISCUSSION

MUC4, located on chromosome 3 in the q29 region [31], encodes a human epithelial mucin that is detected in various epithelial tissues in adult but also in poorly differentiated cells in embryo and fetus [32,33]. Thus MUC4 is expressed early in the primitive gut, before respiratory and digestive epithelial cells have acquired their tissue and cell specificity. Moreover, abnormal expression of MUC4 has been reported in various cancers, such as in pancreatic [34,35] and colon carcinomas [36]. These observations suggest that several distinct functions might be fulfilled by this mucin. To approach these functions, we have determined the complete sequence of *MUC4* cDNA and deduced the peptide organization (Figure 6). Its N-terminus and central sequences have previously been reported [21] as a 27-residue peptide signal, followed by a large domain varying in length from 3285 to 7285 amino acid residues. The C-terminal region of MUC4 shows a very high degree of similarity with the rat heterodimeric glycoprotein complex [22] called SMC. SMC consists of a cell-surface sialomucin (ASGP-1) of 600 kDa associated in a non-covalent manner with an 80 kDa cell-membrane-bound peptide (ASGP-2). Both subunits are translated by a unique cDNA. A GDPH proteolytic cleavage site is present in both peptides in the same position. This suggests that the MUC4 precursor could be cleaved into two subunits and form, as with SMC, a heterodimeric complex. The subunit upstream of the GDPH cleavage site is now called MUC4α and the unit downstream is called MUC4β. Another well-characterized mucin, MUC1, is synthesized on the cell surface as a heterodimeric complex, both subunits originating from a single apomucin precursor [37]. MUC1 is a transmembrane protein [38] for which a soluble form has been reported to be present in cell-culture media and body fluids [39,40]. Although the nucleotide sequence downstream of the 48 bp repeat of MUC4α exhibits similarity with ASGP-1, both peptide domains are different except for regions from amino acid 195 to 284 and from 1252 to 1293. The differences are due to several changes in the reading frame between both apomucins. Thus, a cysteine-rich domain, present in MUC4α, is absent in ASGP-1.

MUC4β subunit is closely related to ASGP-2. Indeed the structural organization of both apomucins is identical and both peptide sequences show more than 60 % similarity. These results suggest that SMC could be considered as the rat homologue of human MUC4. MUC4β is rich in potential N-glycosylation sites, of which 18 out of 21 are conserved within ASGP-2. As in SMC, the hydropathy profile of MUC4β reveals a hydrophobic region of about 24 amino acid residues, which represents a potential membrane-spanning domain. Thus the MUC4 complex, like SMC, is probably also a heterodimeric membrane-associated

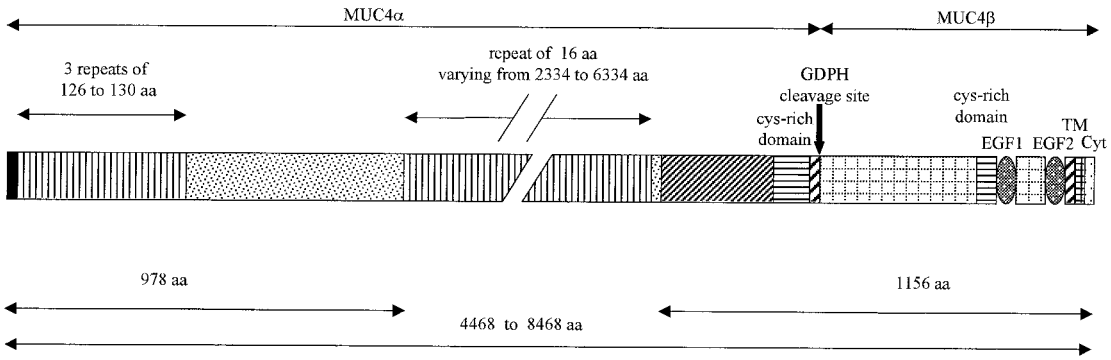

Figure 6 Schematic representation of the structure of MUC4

Black, peptide signal; vertical lines, TR; dense dots, serine, threonine-rich non-repetitive sequence domain; diagonal lines, unique sequence; horizontal lines, cysteine-rich domain; dotted grid, domain rich in potential N-glycosylation sites; hatched ovals, EGF-like domain; TM, potential transmembrane sequence; Cyt, potential cytoplasmic tail.

mucin with its N-terminus orientated extracellularly. SMC is present in different rat tissues as both soluble and membrane-associated forms. For instance, SMC is expressed as two isoforms in the mammary gland and milk; a membrane-associated form (75%) and a soluble or secreted form (25%) [41]. Since no evidence of alternative splicing had been observed, a proteolytic cleavage event was suggested to be responsible for the generation of the soluble form. Our knowledge of MUC4 expression in absorptive and ciliated cells as well as goblet cells suggests that MUC4 could exist as both membrane-associated and secreted forms.

Similar EGF-like domains are found in MUC4β and in ASGP-2. The ASGP-2 EGF1 is considered to interact with the tyrosine kinase p185^{neu}, which is the rat homologue of the proto-oncogene c-ErbB2. ASGP-2 and p185^{neu} are co-immunoprecipitable from cell-surface fractions, and a complex of ASGP-2 and p185^{neu} extracellular domains is formed and secreted from insect cells when the two are co-infected [25]. p185^{neu} shows similarity with the EGF receptor [42], but does not bind EGF. No other p185^{neu} real ligand has been reported. ErbB2 is a member of the class-I EGF receptor tyrosine kinase family, a family of four members, ErbB1–ErbB4. Lupu and colleagues reported a putative ligand, the gp30, that presumably interacts directly with ErbB2 [43]. However, it was not proven that the activity corresponds to a direct ErbB2 ligand. Even without a ligand of its own, ErbB2 can undergo activation by heterologous ligands. EGF and Neu differentiation factor (NDF or heregulin) have been shown to activate the phosphorylation of ErbB2 through the formation of heterodimers, respectively, between ErbB1 and ErbB3 or ErbB4 [44,45]. The ErbB2 gene product is overexpressed in many human cancers, including colorectal [46], non-small-cell lung [47], ovarian [48], breast [49] and uterine cervix carcinoma [50]. It is also expressed in a tissue- and developmental-stage-specific manner [51]. It turned out that the expression pattern of MUC4 is very similar to that of ErbB2 [32,33,47]. In non-small-cell lung cancer, sialomucin expression is associated with ErbB2 overexpression [47]. The heterodimeric membrane-associated isoform of MUC4 could be (as is the case with SMC for p185^{neu}) the natural ligand of the proto-oncogene c-ErbB2. Regulation of ErbB2 receptor activity appears to be very complex. The formation of a MUC4/ErbB2 complex or ErbB2/ErbB1, ErbB2/ErbB3 and ErbB2/ErbB4 may serve to diversify the nature of the intracellular signal elicited by ErbB2. Thus, MUC4 may be a heterodimeric bifunctional cell-surface glycoprotein complex. Recently, MUC1 has been described as a bifunctional cell-surface glycoprotein too [52]. The MUC4 complex, which is very rich in potential O- and N- glycosylation sites, has an extended structure. According to Jentoft [53], the glycosylated polypeptide of 20 amino acid residues is approximately 5 nm long. The MUC4 TR domain varies from 2334 to 6334 residues, so the size of the extended apomucin MUC4 complex varies from 4468 to 8468 residues. This means that MUC4 extends at least 1.12–2.12 μm above the cell membrane, far above all other membrane-associated proteins. For instance, MUC1, which is considered as the largest membrane-associated glycoprotein, extends from 200 to 500 nm [53]. With such size and its putative bifunctionality, MUC4 could be considered as an essential cell membrane-associated glycoprotein, involved in cell–cell communication and the adhesion cascade. Like SMC for p185^{neu}, the MUC4 complex could be involved in a signalling pathway that is required for proliferation and differentiation of epithelial cells.

This work was supported by l'Association de Recherche contre le Cancer and by le Comité du Nord de la Ligue contre le Cancer. N.M. is a recipient of l'Association de Recherche contre le Cancer. We gratefully acknowledge P. Mathon, M. Crépin and C. Mouton for performing automatic sequences, and A. Leclercq and C. Mouton for performing polymorphism analysis. We thank the members of our E.U. consortium CEEBMH4-CT98-3222 for stimulating discussion.

REFERENCES

1 Guzman, K., Bader, T. and Nettesheim, P. (1996) Am. J. Physiol. **270**, L846–L853
2 Braga, V. M. M., Pemberton, L. F., Duhig, T. and Gendler, S. J. (1992) Development **115**, 427–437
3 Lan, M. S., Batra, S. K., Qi, W. N., Metzgar, R. S. and Hollingworth, M. A. (1990) J. Biol. Chem. **265**, 15294–15299
4 Gum, Jr., J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1994) J. Biol. Chem. **269**, 2440–2446
5 Gum, J. R., Hicks, J. W., Swallow, D. M., Lagace, R. E., Byrd, J. C., Lamport, D. T. A., Siddiki, B. and Kim, Y. S. (1990) Biochem. Biophys. Res. Commun. **171**, 407–415
6 Porchet, N., Nguyen, V. C., Dufossé, J., Audié, J. P., Guyonnet Dupérat, V., Gross, M. S., Denis, C., Degand, P., Berheim, A. and Aubert, J. P. (1991) Biochem. Biophys. Res. Commun. **175**, 414–422
7 Dufossé, J., Porchet, N., Audié, J. P., Guyonnet Dupérat, V., Laine, A., Van Seuningen, I., Marrakchi, S., Degand, P. and Aubert, J. P. (1993) Biochem. J. **293**, 329–337
8 Aubert, J. P., Porchet, N., Crépin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D. and Degand, P. (1991) Am. J. Respir. Cell. Mol. Biol. **5**, 178–185
9 Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W. M., Gum, E., Hicks, J. W., Gum, J. R., Byrd, J. C., Siddiki, B. and Kim, Y. S. (1993) J. Biol. Chem. **268**, 5879–5885
10 Bobek, L. A., Liu, J., Sait, S. N. J., Shows, T. B., Bobek, Y. A. and Levine, M. J. (1996) Genomics **31**, 277–282
11 Shankar, V., Pichan, P., Eddy, Jr., R. L., Tonk, V., Nowak, N., Sait, S. N. J., Shows, T. B., Schultz, R. E., Gotway, G., Elkins, R. C., Gilmore, M. S. and Sachdev, G. P. (1997) Am. J. Respir. Cell. Mol. Biol. **16**, 232–241
12 Pigny, P., Guyonnet Dupérat, V., Hill, A., Pratt, W. S., Galiègue-Zouitinia, S., Collyn d'Hooge, M., Laine, A., Van Seuningen, I., Gum, J. R., Kim, Y. S., Swallow, D. M., Aubert, J. P. and Porchet, N. (1996) Genomics **38**, 340–352
13 Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P., Degand, P. and Laine, A. (1998) J. Mol. Evol. **46**, 102–106
14 Ho, S. B., Niehans, G. A., Lyftogt, C., Yan, P. S., Cherwitz, D. L., Gum, E. T., Dahira, R. and Kim, Y. S. (1993) Cancer Res. **53**, 641–651
15 Audié, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B. and Aubert, J. P. (1993) J. Histochem. Cytochem. **43**, 1479–1485
16 Gum, Jr., J. R., Ho, J. J. L., Pratt, W. S., Hicks, J. W., Hill, A. S., Vinall, L. E., Roberton, A. M., Swallow, D. M. and Kim, Y. S. (1997) J. Biol. Chem. **272**, 26678–26686
17 Shekels, L. L., Hunninghake, D. A., Tisdales, A. S., Gipson, I. K., Kieliszewski, M., Kozak, C. A. and Ho, S. B. (1998) Biochem. J. **330**, 1301–1308
18 Khatri, I. A., Forstner, G. G. and Forstner, J. F. (1997) Biochim. Biophys. Acta **1326**, 7–11
19 Gendler, S. J., Lancaster, C. A., Taylor Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalami, E. N. and Wilson, D. (1990) J. Biol. Chem. **265**, 15286–15293
20 Audié, J. P., Tétaert, D., Pigny, P., Buisine, M. P., Janin, A., Aubert, J. P., Porchet, N. and Boersma, A. (1995) Hum. Reprod. **10**, 98–102
21 Nollet, S., Moniaux, N., Maury, J., Petitprez, D., Degand, P., Laine, A., Porchet, N. and Aubert, J. P. (1998) Biochem. J. **332**, 739–748
22 Sherblom, A. P. and Carraway, K. L. (1980) J. Biol. Chem. **255**, 12051–12059
23 Sherblom, A. P., Buck, R. L. and Carraway, K. L. (1980) J. Biol. Chem. **255**, 783–790
24 Sheng, Z., Wu, K., Carraway, K. L. and Fregien, N. (1992) J. Biol. Chem. **267**, 16341–16346
25 Carraway, K. L., Carraway, C. A. C. and Carraway, III, K. L. (1997) J. Mammary Gland Biol. Neoplasia **2**, 187–198
26 Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) Biochemistry **18**, 5294–5299
27 Debailleul, V., Laine, A., Huet, G., Mathon, P., Collyn d'Hooghe, M., Aubert, J. P. and Porchet, N. (1998) J. Biol. Chem. **273**, 881–890
28 Wu, K., Fregen, N. and Carraway, K. L. (1994) J. Biol. Chem. **269**, 11950–11955
29 Kyte, J. and Doolittle, R. F. (1982) J. Mol. Biol. **157**, 105–132
30 Vos, H. L., de Vries, Y. and Hilkens, J. (1991) Biochem. Biophys. Res. Commun. **181**, 121–130
31 Gross, M. S., Guyonnet Dupérat, V., Porchet, N., Bernheim, A., Aubert, J. P. and Van Cong, N. (1992) Ann. Hum. Genet. **35**, 21–26
32 Buisine, M. P., Devisme, L., Savidge, T. C., Gespach, C., Gosselin, B., Porchet, N. and Aubert, J. P. (1998) Gut **43**, 519–524
33 Buisine, M. P., Devisme, L., Copin, M. C., Durand-Réville, M., Gosselin, B., Aubert, J. P. and Porchet, N. (1999) Am. J. Respir. Cell. Mol. Biol. **19**, in the press

34 Balagué, C., Gambus, G., Carrato, C., Porchet, N., Aubert, J. P., Kim, Y. S. and Real, F. X. (1994) Gastroenterology **106**, 1054–1061
35 Balagué, C., Audié, J. P., Porchet, N. and Real, F. X. (1995) Gastroenterology **109**, 953–964
36 Ogata, S., Uehara, H., Chen, A. and Itzkowitz, S. H. (1992) Cancer Res. **52**, 5971–5978
37 Ligtenberg, M. J. L., Kruijshaar, L., Biujs, F., van Meijer, M., Litvinov, S. V. and Hilkens, J. (1992) J. Biol. Chem. **267**, 6171–6177
38 Pemberton, L., Taylor-Papadimitriou, J. and Gendler, S. J. (1992) Biochem. Biophys. Res. Commun. **185**, 167–175
39 Boshell, M., Lalani, E.-N., Pemberton, L., Burchell, J., Gendler, S. J. and Taylor-Papadimitriou, J. (1992) Biochem. Biophys. Res. Commun. **185**, 1–8
40 Burchell, J., Wang, D. and Taylor-Papadimitriou, J. (1984) Int. J. Cancer **34**, 763–768
41 Rossi, E. A., McNeer, R. R., Price-Schiavi, S. A., Van den Brande, J. M. H., Komatsu, M., Thompson, J. F., Carraway, C. A. C., Friegien, N. L. and Carraway, III, K. L. (1996) J. Biol. Chem. **271**, 33476–33485
42 Gullick, W. J. (1990) Int. J. Cancer suppl. **5**, 55–61
43 Lupu, R., Colomer, R., Zugmaier, G., Sarup, J., Slamon, D. and Lippman, M. E. (1990) Science **249**, 1552–1555
44 Wada, T., Qian, X. and Greene, M. I. (1990) Cell **61**, 1339–1347
45 Carraway, III, K. L. and Cantley, L. C. (1994) Cell **78**, 5–8
46 Kapitanovic, S., Radosevic, S., Kapitanovic, M., Andelinovic, S., Frerencic, Z., Tavassoli, M., Primorac, D., Sonicki, Z., Spaventi, S., Pavelic, K. and Spaventi, R. (1997) Gastroenterology **112**, 1103–1113
47 Yu, C.-J., Shun, C.-T., Yang, P.-C., Lee, Y.-C., Shew, J.-Y., Kuo, S.-H. and Luh, K.-T. (1997) Am. J. Respir. Crit. Care Med. **155**, 1419–1427
48 Meden, H. and Kuhn, W. (1997) Eur. J. Obstet. Gynecol. Reprod. Biol. **71**, 173–179
49 Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., Levin, W. J., Stuart, S. G., Udove, J., Ullrich, A. and Press, M. F. (1989) Science **244**, 707–712
50 Costa, M. J., Walls, J. and Trelford, J. D. (1995) Am. J. Clin. Pathol. **104**, 634–642
51 Kokai, Y., Cohen, J. A., Drebin, J. A. and Greene, M. I. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 8498–8501
52 Mockensturm-Gardner, M., Rowles, J. and Gendler, S. J. (1998) 5th International Workshop on Carcinoma-Associated Mucin, Abstract, D6
53 Jentoft, N. (1990) Trends Biochem. Sci. **15**, 291–294

Received 14 September 1998/11 November 1998; accepted 10 December 1998