Human mucin gene *MUC4*: organization of its 5'-region and polymorphism of its central tandem repeat array

Sverine NOLLET^{*1}, Nicolas MONIAUX^{*1}, Jacques MAURY^{*}, Danile PETITPREZ^{*}, Pierre DEGAND^{*†}, Anne LAINE^{*}, Nicole PORCHET^{*†} and Jean-Pierre AUBERT^{*†2}

^{*}INSERM Unit 377, Place de Verdun, 59045 Lille Cedex, France, and [†]Laboratoire de Biochimie et de Biologie Molculaire, l'Hpital C. Huriez, Centre Hospitalier Rgional et Universitaire, 59037 Lille Cedex, France

In a previous study we isolated a partial cDNA with a tandem repeat of 48 bp, which allowed us to map a novel human mucin gene named *MUC4* to chromosome 3q29. Here we report the organization and sequence of the 5'-region and its junction with the tandem repeat array of *MUC4*. Analysis of three overlapping genomic clones allowed us to obtain a partial restriction map of *MUC4* and to locate the complete 48 bp tandem repeat domain on a *Pst*I/*Eco*RI genomic fragment that exhibits a very large variation in number of tandem repeats (7–19 kb). cDNA clonal extension allowed us to obtain the entire 5' coding region of

MUC4. Exon 1 consists of a 5' untranslated region and an 82 bp fragment encoding the signal peptide. This latter shows a high degree of similarity to the signal peptide of another apomucin, ASGP-1. Exon 2 is extremely large and contains a unique sequence that is followed by the whole tandem repeat domain. It encodes only one cysteine residue, making *MUC4* different from mucin genes belonging to the 11p15.5 family. Moreover, an intron downstream from the tandem repeat array consists mainly of a 15 bp tandem repeat that exhibits a polymorphism in having a variable number of tandem repeats.

INTRODUCTION

Mucins constitute a complex family of glycoconjugates, characterized by a large amount of O-glycans linked to the peptide backbone (reviewed in [1–3]). These O-linked glycans usually account for more than 50 % of the molecular mass of the mucin. The potential O-glycosylation sites (threonine and serine residues) are located mainly in tandemly repeated amino acid sequences. Mucins can be subdivided into secreted mucins and 'mucin-like' membrane-bound O-glycoproteins. So far nine human mucin genes, designated *MUC1*, 2, 3, 4, 5A, 5B, 6, 7 and 8 in the international nomenclature, have been identified (reviewed in [1]; *MUC8* reviewed in [4]). *MUC1* is located on 1q21–24, *MUC3* on 7q22, *MUC4* on 3q29, *MUC7* on 4q13–21 and *MUC8* on 12q24.3. *MUC2*, *MUC5AC*, *MUC5B* and *MUC6* are clustered on 11p15.5. Several partial cDNA species containing tandem repeats (TRs) have been isolated in our laboratory from a human tracheobronchial cDNA library and correspond to the three distinct genes *MUC4*, *MUC5AC* and *MUC5B* [5–7]. The first partial cDNA from *MUC4*, called JER64, contains 39 repeats of 48 nearly identical nucleotides that encode a Thr/Ser-rich (consensus sequence ATPLPVTDTSSASTGH) potentially glycosylated peptide mucin domain [6]. *MUC4* displays genetic polymorphism, which suggests that it is related to a variable number of tandem repeats (VNTR) specific to each individual [8].

The gene-expression pattern of *MUC4* has been determined by using Northern blotting and/or hybridization *in situ*. *MUC4* mRNA species are expressed in colon, stomach, cervix and lung

[9,10] but are not detected in normal pancreas, gall bladder and breast tissues. The expression of *MUC4* mRNA species is ubiquitous in epithelial cells, where it is equally expressed in goblet and ciliated cells of the trachea and bronchi. Similarly, in the intestinal mucosa *MUC4* is expressed in both goblet and absorptive cells.

Abnormal expression of *MUC4* has been reported in various carcinomas. For example, increased expression of *MUC4* mRNA species is observed in pancreatic carcinoma and cancer cell lines [11,12] and mammary carcinomas [13]. Moreover, the frequent occurrence of increased *MUC4* transcripts in a variety of non-small-cell lung cancers [14] and in colon carcinomas [15] indicates that the overexpression of this gene might have an important, albeit undefined, role in tumour biology. Therefore it is of considerable interest to define more fully the genomic structure of *MUC4*.

Here we report the genomic organization of the 5'-region and the whole TR array of the human mucin gene *MUC4*. The nature of the VNTR polymorphism of the 48 bp TR region is also described.

EXPERIMENTAL

Screening of genomic libraries

The probe designated JER64 used in this study corresponds to the *MUC4* cDNA probe (1.83 kb) and contains 39 identical 48 bp TRs [6]. Human genomic DNA fragments (of approx.

Abbreviations used: endo H, endoglycosidase H; RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcription PCR; TR, tandem repeat; VNTR, variable number of tandem repeats.

¹ The first two authors contributed equally to this work and should therefore be considered as equal first authors.

² To whom correspondence should be addressed at INSERM U-377 (e-mail jpa@lille.inserm.fr).

The nucleotide sequence data reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession numbers AJ000281 and AJ000282.

20 kb) were obtained by partial digestion with *Sau3A* and were cloned into *Bam*HI sites of a λ EMBL4 phage vector. Screening of the library was performed with the JER64 probe. One positive clone (ANT55) with an insert of approx. 15.5 kb was obtained.

To isolate larger genomic clones of *MUC4*, a human placenta genomic cosmid DNA library in pWE15 (Stratagene) was screened with the JER64 probe. Two positive clones (LEA2 and LEA47), each containing inserts of approx. 45 kb, were obtained and analysed. The same cosmid genomic library was also screened with the insert of the RAC3 clone described below. One positive clone with an insert of approx. 40 kb (LEA51) was isolated and studied.

Restriction mapping of cosmids

The restriction mapping strategy of Wahl et al. [16] was modified slightly, as described previously [17].

Study of the 48 bp tandem repeat region

To evaluate the number of 48 bp repeats, we first cut the LEA2 cosmid with *Pst*I and *Eco*RI at positions that flank the region containing these repeats and isolated the fragment. Complete or partial digestion with *Dde*I was achieved with 10 units/ μ g of DNA for 4 h or 0.14 unit/ μ g of DNA for 1 h respectively. Southern blot analysis was conducted with the JER64 probe, as described previously [6].

RNA extraction

Total RNA was extracted from normal human colon mucosa with the guanidine isothiocyanate/CsCl method [18].

cDNA library preparation and screening

Total RNA from human colon mucosa was prepared and used as a template for cDNA synthesis. All details of double-stranded cDNA synthesis and cloning into λ gt11 vector were as described by the commercial supplier (Amersham).

Screening of the cDNA library was performed with the JER64 probe. One single positive clone, JER103, was isolated and sequenced.

Cloning into pKS

The fragments of interest from phage or cosmid clones were subcloned into pBluescript KS(+) vector from Stratagene.

5' Rapid amplification of cDNA ends (RACE) procedures

The 5'/3' RACE kit (Boehringer Mannheim) was used to synthesize first-strand cDNA species from total human colon RNA (2 μ g) with specific primers for *MUC4* (NAU124, NAU155, NAU168 and NAU287; their locations are given below). Terminal transferase was then used to add a poly(dA) tail to the 3' end of the cDNA. RACE-PCR experiments were performed in 50 μ l reaction volumes containing 5 μ l of 10 $\times$ buffer (100 mM Tris/HCl/15 mM MgCl₂/500 mM KCl, pH 8.3), 5 μ l of 10 mM deoxynucleoside triphosphates, 5 μ l of poly(dA)-tailed cDNA, 12.5 pmol of each primer [a set of four specific primers, NAU138, NAU156, NAU169 and NAU 288, and an oligo(dT) anchor primer were used], and 2 units of *Taq* DNA polymerase (Boehringer Mannheim). After being overlaid with 60 μ l of mineral oil (Sigma), the mixture was denatured at 94 °C for 2 min followed by 30 cycles at 94 °C for 1 min, 60 °C for 1 min and finally 72 °C for 2 min. The elongation step was extended for an additional 15 min period. A 1 μ l sample of the primary

amplification product was further amplified by a second PCR reaction with a nested specific primer of *MUC4* (NAU138, NAU156, NAU204, NAU289 and NAU327) and the PCR anchor primer. The thermal cycling protocol used was the same as for the primary RACE amplification step, except that the annealing was performed at 62 °C. PCR experiments were performed with a Perkin-Elmer thermal cycler model 480.

Various antisense primers were used to extend the 5' end sequence: NAU124 (nt 565–588) (see Figure 4), NAU138 (nt 538–564), NAU155 (nt 308–325), NAU156 (nt 275–301), NAU168 (nt 105–124), NAU169 (nt 81–100), NAU204 (nt 51–77), NAU287 (nt –15 to 6), NAU288 (nt –41 to –21), NAU289 (nt –76 to –56) and NAU327 (nt –130 to –109).

Reverse transcription and amplification

Human tracheal poly(A)⁺ RNA (0.5 μ g) (Clontech) and 1 μ g of total RNA extracted from human colon mucosa were reverse transcribed with the 1st-STRAND[®] cDNA synthesis kit (Clontech) with random primers in accordance with the manufacturer's instructions. First-strand cDNA (8 μ l) was amplified by PCR with various primers: NAU162 (sense) (nt 2–19), NAU138 (antisense) (nt 538–564), NAU370 (sense) (nt –71 to –48), NAU224 (antisense) (nt 955–974), NAU174 (sense) (nt 3015–3035) and NAU99 (antisense) (nt 3093–3109). The thermal cycling protocol was the same as that described above.

Cloning of amplification products

RACE-PCR and reverse transcription PCR (RT-PCR) products were separated by electrophoresis, excised and purified with Preps DNA purification resin (Promega), and finally cloned into pGEMT (Promega) or pCR2.1 (Invitrogen) vector.

Plasmid DNA purification

The Wizard[®] minipreps DNA purification system (Promega) was used in accordance with the manufacturer's instructions.

DNA sequencing

Clones were sequenced on both strands by the dideoxynucleotide chain termination method, by using [α -³²S]dATP with Sequenase version 2.0 (U.S. Biochemical Corp.) and synthetic oligonucleotides corresponding to the T7 and T3 primers of the pKS plasmid, and to the T7 and –40 primers of the pGEMT or pCR2.1 vector. Part of the sequence was determined by primer walking with primers specific to *MUC4*. The locations of these specific primers on the cDNA are indicated above (in the section on 5' RACE and reverse transcription procedures) except that of NAU103 (5'-GTAATGCGAATGCACCAGTG-3', antisense) located within an intron. We performed DNA sequencing directly on cosmids, as described previously [17]. The inner regions of some clones were sequenced with exonuclease III-deleted clones. Sequencing reaction mixtures were subjected to electrophoresis on 6% (w/v) polyacrylamide gel (Sequagel-6[®]; National Diagnostics). Nucleic acid and protein sequence results were analysed with PC/GENE Software.

Transcription and translation assays *in vitro*

RT370-224 is an RT-PCR product (NAU370/NAU224) containing the cDNA coding a peptide with a predicted size of 34 kDa. This peptide comprises the 27-residue N-terminal signal sequence and two potential N-glycosylation sites. The corre-

sponding cDNA was cloned into pCR2.1 vector under the control of the T7 promoter. Transcription and translation experiments *in vitro* were performed with the TNT Coupled Reticulocyte Lysate System (Promega) in accordance with the manufacturer's instructions. The amino acid mixture lacking methionine, supplemented with [³⁵S]methionine (approx. 1000 Ci/mmol; Amersham), was used. To detect a signal peptide, canine pancreatic microsomal membranes (Promega) were added to the TNT Coupled Reticulocyte Lysate System for co-translational processing of translation products. To determine the extent of N-glycosylation of the translation products, ³⁵S-labelled peptide was incubated overnight at 37 °C with 10 m-units of endoglycosidase H (endo H; Boehringer Mannheim) in a buffer containing 0.02% SDS and 0.1 M sodium citrate, pH 5.5. Translation products were analysed by SDS/PAGE in slab gels with a 10–30% (w/v) polyacrylamide gradient. Rainbow[™] coloured protein molecular mass markers (Amersham) varying in size from 14.3 to 220 kDa were used to determine the molecular mass.

Southern blot analysis

Human genomic DNA from 18 healthy volunteers was digested with *Pst*I and *Eco*RI restriction endonucleases; 12 out of the 18 DNA species were digested with *Dde*I (complete digestion). Fragments were separated by electrophoresis and transferred to nylon N⁺ membrane (Amersham). They were hybridized with three different probes and washed as described previously [6].

RESULTS

MUC4 genomic clones

We isolated three genomic clones from *MUC4* with the JER64 probe. One of these, designated ANT55, was isolated from a phage library and the other two, LEA2 and LEA47, from a cosmid library. Alignment and partial restriction maps of these three clones are shown in Figure 1. The two cosmid clones overlap ANT55, which is 15.5 kb in length. A 2.5 kb *Eco*RI fragment located at the 3' end of ANT55 was subcloned into pKS vector. This fragment, designated HOR1, was completely sequenced. It consists of 109 repetitions of very similar 15 bp units (consensus sequence GGTGTGGAAGGTATG) and two unique flanking sequences of 795 and 62 bp at the 5' and 3' ends respectively. The 15 bp TR domain of HOR1 has been submitted to the EMBL Data Bank with accession number AJ000282. The oligonucleotide primer NAU103 situated on the 5' end of HOR1 (its position is shown in Figure 1) was chosen and used to perform direct sequencing on ANT55 clone. The 48 bp TR domain of JER64 was located 85 bp upstream of the *Eco*RI restriction site. The direction of the transcription could then be determined as shown in Figure 1. The primer NAU103 was also used to perform direct sequencing on LEA2 and LEA47 cosmid clones. The same sequence was determined. The region of HOR1 containing only the 15 bp repeats did not hybridize to RNA species from human colon and bronchus mucosae by Northern blot or *in situ* hybridization analyses (results not shown).

The 48 bp tandem repeat region

*Pst*I/*Eco*RI digestion allowed us to isolate the smallest fragment containing the whole 48 bp repetitive domain. In the ANT55 insert this fragment was 11 kb in length; it was approx. 20 kb in length in the LEA2 and LEA47 inserts. Only a few usual restriction enzymes were able to cut within this 48 bp repetitive sequence. Among these, the *Rsa*I restriction endonuclease cleaved

the LEA2 insert into four distinct bands (approx. 8, 5, 3.2 and 2 kb), which hybridized with the JER64 probe. Moreover, a combined enzyme digestion *Pst*I/*Eco*RI/*Rsa*I gave the same bands. So the sum of the sizes of the subfragments included into the *Pst*I/*Eco*RI fragment of the LEA2 insert allowed us to calculate a more precise size for the whole *Pst*I/*Eco*RI fragment, which was predicted to be approx. 18.2 kb. This region was completely devoid of all conventional endonuclease restriction sites and could thus not be cloned into a plasmid vector, thereby making direct sequence determination impossible. Some restriction sites that are not present in the multicloning site of the vector exist in this region (*Dde*I, *Fok*I, *Hph*I, *Ksp*632I, *Mae*III, *Mbo*II, *Mn*II, *Sfa*NI) but they cut it into a range of very small fragments (results not shown). However, because one *Dde*I restriction site (CTNAG) exists in most of the JER64 48 bp repeats (in 36 out of 39 units), *Dde*I was used to cut this region. Figure 2 shows the 18.2 kb *Pst*I/*Eco*RI fragment of LEA2 digested to completion with the restriction enzyme *Dde*I. Only two ethidium bromide-stained fragments (48 and 96 bp) were observed, as expected from the JER64 nucleotide sequence. These two bands hybridized with the JER64 probe. The presence of a 96 bp fragment is due to two adjacent units lacking one *Dde*I site. The fact that no additional band was observed strongly suggests that the TR region of the LEA2 insert is not interrupted by an additional unique sequence. Partial *Dde*I digest of the 18.2 kb *Pst*I/*Eco*RI fragment of LEA2 (Figure 2) showed a ladder where individual bands consisted of 48 bp multiples. Beyond approx. 20 bands, a smear was observed and precise estimation of the number of units was difficult. The same observation was made with ANT55 and LEA47. Thus the TR arrays in the phage clone and in the cosmid clones appear to contain approx. 208 and 380 uninterrupted individual TR units respectively. Southern blot analysis of the *Dde*I fragments from DNA species isolated from lymphocytes of 12 individuals gave the same result: only fragments of 48 and 96 bp hybridized to the JER64 probe (results not shown).

JER103 cDNA clone

The JER64 cDNA probe was also used to screen a human colon mucosa cDNA library. One positive clone was obtained and designated JER103 (Figure 3). The JER103 insert consists of 2653 bp, of which 200 bp belong to the 48 bp TR recognized by JER64. A *Pst*I restriction site was found 263 bp upstream of the TR (labelled 3 in Figure 1). The choice of the reading frame was directed by that of the JER64 cDNA clone. The nucleotide sequence of JER103 revealed the presence of a unique fragment encoding a threonine/serine-rich peptide. This sequence was found immediately upstream of the 48 bp repeat array on the cosmid inserts, indicating that both of these unique and repetitive Thr/Ser-rich regions are encoded by a single exon. The JER103 insert sequence is identical with that found in LEA2 and LEA47 cosmid and ANT55 phage clones, this last being 439 bp shorter.

Extension of the JER103 5' end sequence

Two antisense primers, NAU124 and NAU138, were used to extend the sequence by 5' RACE on human colon mucosa total RNA. One 358 bp cDNA fragment was obtained, cloned and designated RAC1 (Figure 3). Its 3'-end sequence overlaps the 5' end sequence of JER103, indicating that it contains the JER103 5'-extended end. This sequence is identical with that found in LEA2 and LEA47 cosmid clones immediately upstream of the sequence corresponding to that of the JER103 clone. This experiment was followed by a second 5' RACE procedure with two novel primers deduced from the sequence of the RAC1 clone

Figure 1 Partial restriction map of three genomic clones (ANT55, LEA2 and LEA47) and one cDNA clone (JER103) from *MUC4*

The vertically hatched box represents the 48 bp TR array. The diagonally hatched box represents the unique sequence encoding a threonine/serine-rich region. The *PstI*/*EcoRI* fragment that delimited the 48 bp TR region is 11 kb in length for the phage clone and approx. 18.2 kb in length for the cosmid clones. Primers and their directions are indicated (not to scale) by horizontal arrows and their NAU numbers (their locations are given in the Experimental section). All genomic inserts are flanked by *EcoRI* sites of the vectors indicated by vertical arrows. *PstI* restriction sites used in polymorphism analysis are numbered in a circle.

(NAU155 and NAU156). This produced a 300 bp fragment (RAC2). The sequence of the 3' region of the RAC2 clone is the same as that of a 219 bp stretch found in the three genomic clones ANT55, LEA2 and LEA47. However, the first 81 bp of the RAC2 clone are not found within these genomic clones. Thus an intronic region was suspected. Another 5' RACE procedure with the oligonucleotides NAU168, NAU169 and NAU204 (chosen in the sequence of the RAC2 clone) produced a fragment of 183 bp (RAC3). Two other RAC clones, RAC4 (140 bp) and RAC5 (352 bp), were then obtained by using four novel oligonucleotides (NAU287, NAU288, NAU289 and NAU327). These

five RAC clones allowed us to extend the 5' end sequence of JER103 over 942 bp.

The RAC3 probe does not hybridize to Southern blots from LEA2 and LEA47 DNA species, showing that the 5' sequence of 542 bp obtained by compiling the sequences of clones RAC3, RAC4 and RAC5 is not located on these cosmids. This indicates that this coding sequence is situated at least 15 kb upstream of the RAC1–2 sequence and that the 15 kb fragment constituting the 5' end of LEA47 corresponds to a large intron (Figure 3).

To obtain the 5'-region of *MUC4*, the pWE15 cosmid genomic library was screened with probe RAC3. One novel clone

Figure 2 Characterization of the 48 bp repeat in the LEA2 clone

After electrophoretic separation in an agarose gel, blot analysis was conducted with the JER64 probe. Lanes A–C, LEA2 was cut at *Pst*I and *Eco*RI sites flanking the 48 bp TR region: lane A, *Pst*I/*Eco*RI fragment; lane B, *Pst*I/*Eco*RI fragment partly digested for 1 h with *Dde*I (0.14 unit/ μ g of DNA); lane C, *Pst*I/*Eco*RI fragment totally digested with *Dde*I (10 units/ μ g of DNA). Lane D, JER64 insert totally digested with *Dde*I.

designated LEA51 was isolated and its partial restriction map revealed one 2.2 kb *Eco*RI fragment situated at the 3' end of this clone. This fragment was subcloned in pKS vector and partly sequenced. Its 3' part consists of an intronic region of 400 bp. The 542 bp upstream of this intronic region showed 100% sequence similarity to the 542 bp compiled cDNA sequences of RAC3–5.

The exon–intron boundaries were determined by comparing the genomic and cDNA nucleotide sequences. Splice acceptor and donor sequences agree with the 'GT–AG' rule proposed by Mount [19]. This intron is class 1 because it interrupts the coding sequence between the first and second bases of the codon [20].

RT–PCRs with the primers NAU162 and NAU138 were performed starting from human RNA species from tracheo-bronchial and colon mucosae; 563 bp amplification products were generated and cloned. Their nucleotide sequence showed 100% similarity to the corresponding sequences obtained by compiling RAC2 and RAC1 sequences.

Analysis of the nucleotide and deduced amino acid sequences of the MUC4 5' end cDNA

The compiled nucleotide sequences of the different cDNA clones obtained allowed us to establish the whole coding sequence of the MUC4 cDNA 5' part and its junction with the expansive 48 bp TR region. The 460 bp 5' untranslated region is followed by a region encoding an open reading frame of 978 residues

(Figure 4). This open reading frame is continuous with that of the 16-residue TR.

A methionine residue starting at nt 1 is contained within the optimal context for initiation of translation, GCCGCAGC-CATGA, as described by Kozak [21], except for the –4 and +4 positions, which are A instead of C and G respectively. The Kyte–Doolittle [22] hydropathy plot of the first 200 residues of the deduced sequence showed that the initial 27 residues encoded by exon 1 are very hydrophobic, suggesting that these comprise the putative signal peptide. Three potential cleavage sites exist between amino acids 23 and 24, 27 and 28, and 28 and 29. To confirm that this peptide can act as a signal sequence, the RT370-224 clone (1045 bp) [see its position in Figures 3(C) and 4] was analysed by transcription and translation *in vitro* in the presence or absence of dog pancreatic microsomes (Figure 5). Lane 1 (in the presence of microsomes) shows three bands, where the upper is stronger and corresponds to a translated product with an apparent molecular mass of 43 kDa. Lane 2 (after treatment with endo H) shows two bands, where the smaller (corresponding to a 39 kDa product) migrates with the same mobility as the third band before treatment with endo H and the weaker like the second. The treatment with endo H in lane 2 was probably incomplete. The upper band in lane 1 is probably due to two *N*-glycans, the second to one *N*-glycan and the very weak third to the non-glycosylated one. In the absence of microsomes (lane 3), the translated product (324 residues) had an apparent molecular mass of 40 kDa. Thus the translated product of lane 2 is smaller than the 40 kDa peptide formed in the absence of microsomes, indicating that it has been cleaved as a result of translation in the presence of the membranes containing signal peptidase. Searching the GenBank database, we noticed that a high degree of similarity exists between signal peptides of MUC4 and rat ASGP-1 [23], as well as between MUC1 [24] and mouse Muc1 [25], and MUC2 [26] and rat Muc2 [27] (Figure 6A). A 62% similarity between rat ASGP-1 and MUC4 was observed at the nucleotide level, whereas a 59% similarity was seen at the protein level. Similarity between the two signal peptides is particularly striking when considering their respective C-terminal regions, where 12 residues out of 15 are perfectly conserved (80% similarity) (Figure 6B).

The region from nt 83 to nt 2934, together with the TR array, forms a single large exon. This region is found to encode a unique 951-residue sequence typical of apomucins, comprising 21.5% threonine and 19% serine. The sequence TXXP, considered to be a major O-glycosylation site [28], is repeated on 21 occasions. This region is also proline-rich (7.4%). Three potential N-glycosylation sites are also found in this sequence at positions 235, 260 and 622 (it is likely that at least two were glycosylated *in vitro*) and only one cysteine residue at position 256. The sequence for this unique Thr/Ser-rich region bears no significant similarity to the 48 bp TR of JER64 or to any other sequences from the GenBank database. Closer examination reveals that this sequence contains three subregions that share a high degree of similarity. These three repeated units begin at residue 43. The first and the second repeats are both 126 residues in length, whereas the third is 130 residues long. The amino acid composition of each subregion contains 24% threonine and 20% serine and is typical of apomucins. A tetrabasic amino acid RKRR site is also found at position 822.

Determination of intron–exon boundaries on the 3' end of the 48 bp TR region

To determine the 3' end of the central exon, RT–PCR experiments were performed with NAU174 and NAU99 as primers (see their positions in Figure 1), starting from several sources of poly(A)⁺

Figure 3 Organization of the 5'-terminal region of *MUC4*

(A) Schematic representation of the exon/intron structure of the 3' end of clone LEA51 and the 5' end of clone LEA2. The open box on the left represents the 5' untranslated region, the black box the region encoding the signal peptide, the diagonally hatched box the unique sequence encoding a threonine/serine-rich region and the vertically hatched box on the right the 5'-terminal region of the *MUC4* 48 bp TR array. The broken line represents the intronic region. (B) Illustration of the compiled sequence of the different cDNA clones obtained. The primers and their directions are indicated by horizontal arrows and their NAU numbers (their locations are given in the Experimental section). (C) Location and length of the JER103 cDNA clone and of cDNA clones obtained by 5' RACE and RT-PCR experiments.

RNA (tracheal mucosa or colon mucosa). The different amplification products were subcloned into pCR2.1 vector and sequenced. The clones were designated RT174-99. The sequences obtained (Figure 4) were analysed and compared with those obtained from HOR1 (the 2.5 kb *EcoRI* fragment of ANT55) (Figure 1) and its 5' extension with NAU103 as primer with direct sequencing. A short intron of 333 bp whose location is marked in Figure 4 is 37 bp downstream of the 48 bp repeat, as evidenced when comparing cDNA and genomic sequences. Splice acceptor and donor sequences conform to the 'GT-AG' rule. This intron might also be categorized as class 1.

Polymorphism studies

Lymphocyte-derived DNA isolated from 18 volunteers (unrelated Caucasian individuals from the north of France) was

digested with *PstI* and *EcoRI*. As shown previously, the 48 bp TR array is flanked by a *PstI* site (labelled 3 on Figure 1) and an *EcoRI* site, and the 15 bp TR array is flanked by the same *EcoRI* site and a *PstI* site (labelled 4 on Figure 1). Three additional *PstI* sites permitted the cleavage of the unique 5'-exonic sequence close to the 48 bp TR domain into two fragments. This double digestion was therefore useful in investigating contiguous fragments for polymorphism studies of *MUC4* (Figure 7). The two *PstI* fragments located on the 5' part of *MUC4*, which hybridized with the JER103 probe, were constantly 2 and 0.5 kb in length in all individuals tested. In contrast, the two TR arrays showed high degrees of polymorphism. With each probe, Southern blot analysis demonstrated two bands with the same intensity in 15 out of 18 individuals. These two bands were considered to be allelic forms of each TR domain; no additional band indicated the presence of *PstI* or *EcoRI* sites within the TR domains. In the

Figure 4 Compiled nucleotide sequences and deduced amino acid sequences of the RAC1-5, JER103 and RT174-99 cDNA clones

Nucleotide positions are indicated by the numbers at the left, and amino acid positions at the right. The locations of primers NAU370 and NAU224 are indicated by black arrows. The secretory protein signal sequence is underlined. The asterisks indicate the positions of cysteine residues (lines 20 and 260). Black downwards-pointing arrowheads indicate the positions of the introns. Open arrows indicate the three repeat units found at the beginning of the Thr/Ser-rich region. Three putative N-glycosylation sites are boxed. A tetrabasic amino acid sequence RKRK is doubly underlined. The sequence is presented with only one complete 48 bp tandem repeat. The number of repeats (*n*) varies from approx. 145 to approx. 395 in the present study. The cDNA sequence reported in this figure has been submitted to the EMBL Data Bank with accession number AJ000281.

Figure 5 [³⁵S]Methionine labelling of the expression product of RT370-224 synthesized in a cell-free system in the absence and in the presence of microsomes and of endo H

Transcription and translation reactions with the TNT reticulocyte lysate system were performed and the reaction products were processed as described in the Experimental section. Lane 1, incubation with 1 μ g of RT370-224 and microsomes; lane 2, incubation with 1 μ g of RT370-224, microsomes and endo H; lane 3, incubation with 1 μ g of RT370-224; lane 4, incubation with 1 μ g of pCR2.1 vector; lane 5, incubation with 1 μ g of pCR2.1 vector and microsomes. The sizes of the Rainbow[®] coloured protein molecular mass markers are indicated at the right.

48 bp TR domain, 13 distinct alleles were observed from 19 to 7 kb (A, 19 kb; B, 17 kb; C, 14 kb; E, 12 kb; F, 11.5 kb; G, 11 kb; H, 10.5 kb; I, 10 kb; J, 9.2 kb; K, 8.9 kb; L, 8.5 kb; M, 7.5 kb; N, 7 kb). The largest allele (A) was the most common. However, most individuals exhibiting the largest allele (12 out of 18) were heterozygous (10 out of 18), as only two displayed homozygosity. In these individuals, the size of the second allele varied from 12 kb (allele E) to 7 kb (allele N). Six individuals lacking the common allele A exhibited various other alleles (B, C, E, G, H, I and L); one of these was homozygous for E. The length of the *Pst*I/*Eco*RI fragment containing the 48 bp TR domain in LEA2 and LEA47 corresponds to allele B. This in ANT55 was like allele G. In the intronic 15 bp TR domain of *MUC4*, 12 distinct alleles were observed from 6.5 to 1.5 kb (a, 6.5 kb; b, 6.2 kb; c, 5.5 kb; d, 5.2 kb; e, 5.0 kb; f, 4.0 kb; h, 3.6 kb; i, 3.3 kb; j, 2.8 kb; k, 2.4 kb; l, 1.8 kb; m, 1.5 kb). The most common allele found in 12 out of 18 individuals was h. Ten individuals were heterozygous and two homozygous; these two

Figure 6 Comparison of several mucin signal peptides

(A) Alignment of the deduced amino acid sequences of the signal peptide of MUC1 with that of its mouse homologue, the signal peptide of MUC2 with that of its rat homologue, and the signal peptide of MUC4 with that of rat ASGP-1. Dashes indicate gaps introduced into the sequence for alignment purposes. Identical amino acids are boxed. The numbers indicate amino acid positions. (B) Nucleotide sequence similarity between the signal sequence of MUC4 and that of rat ASGP-1. Identical nucleotides are boxed. The numbers indicate nucleotide positions.

Figure 7 Location of sequence polymorphisms

Genomic DNA prepared from 18 random unrelated individuals was digested to completion with *PstI* and *EcoRI*. The same blot of *PstI/EcoRI*-digested DNA samples was hybridized sequentially with: *PstI*¹-*PstI*² 5' probe (A); *PstI*²-*PstI*³ 5' probe (B); MUC4 TR probe (JER64) (C); and 3' probe (HOR1) (D).

individuals were also homozygous for A. The *EcoRI/PstI* fragment containing the 15 bp TR domain in the two cosmids (LEA2 and LEA47) corresponds to allele h and that of the phage clone (ANT55) to allele j.

DISCUSSION

A human tracheobronchial mucin cDNA from a mucin gene named *MUC4* has previously been isolated in our laboratory [6]. The 1.83 kb JER64 clone insert consists entirely of nearly identical 48 bp repeats, encoding a Thr/Ser-rich stretch. This cDNA was mapped to chromosome 3q29 [8]. The whole genomic TR domain hybridizing to the JER64 probe is defined by a *PstI/EcoRI* band of approx. 18.2 kb in cosmid clones and 11 kb in phage clones. Digestion with *DdeI* allowed us to establish that the *PstI/EcoRI* fragment consists entirely of 48 bp repeats because no fragments larger than 96 bp (corresponding to two units) were observed. Thus the TR array in cosmid clones seems to contain approx. 380 uninterrupted individual units of 48 bp (approx. 18.2 kb) that are almost identical. Moreover, Southern blot analysis confirmed that the *MUC4* TR array of 12 individuals consists only of 48 bp units because alleles cut to completion by restriction enzyme *DdeI* showed only 48 or 96 bp units. Therefore the TR region of *MUC4* is uninterrupted and located within a single exon.

Gross et al. [8] suggested that the polymorphism observed between individuals with the JER64 probe was VNTR, owing to the repetitive structure of this cDNA. However, the restriction sites for the enzymes used in this study did not accurately flank the 48 bp TR. In the present study we chose to use *PstI* and *EcoRI*, which just flank the 48 bp TR array, thereby enabling the isolation of the whole TR domain. Eighteen unrelated individuals were examined by Southern blot analysis; 13 different alleles were detected. In contrast, the unique 5' sequences flanking the 48 bp TRs are of the same length in all genomic or cDNA clones analysed. Thus *MUC4* exhibits a length polymorphism in its TR array that can be characterized as the VNTR type. Alleles observed vary between 7 and 19 kb and correspond to a variation in the number of 48 bp TRs, ranging from approx. 145 to approx. 395 units. Nevertheless, all the individuals studied had at least one large allele. A similar expansive variation in the number of repeats has also been demonstrated for *MUC1*, whose alleles seem to contain between 20 and 125 TRs [24]. Carvalho et al. [29] have reported that individuals with small *MUC1* genotypes are more susceptible to developing gastric carcinoma, suggesting that these mucin gene VNTR variations have a possible functional significance.

The human mucin genes are characterized by a large TR array coding for a peptide that is typically rich in hydroxylated amino acid residues. The largest 60 bp TR region of *MUC1* is 7.5 kb [24]; that of *MUC2* is approx. 8 kb, corresponding to approx. 115 individual TR units of 69 bp [30]. The central region of *MUC7* consists of six highly similar TRs of 69 bp [31]. As with *MUC1*, *MUC2* and *MUC7*, the TR region of *MUC4* is not interrupted by a unique segment. In contrast, Desseyn et al. [17] have indicated that the irregular repeat of 87 bp of *MUC5B* is interrupted by seven unique conserved subdomains that are cysteine-rich. The central exon of *MUC5B* has been completely sequenced and encompasses 10713 bp. The sequence of the TR domain of *MUC5AC* is still incomplete, but in this gene, as in *MUC5B*, this domain is interrupted several times by cysteine-rich unique subdomains [7]. Only partial information is available about the size of the repeat region of *MUC3*, *MUC5AC* and *MUC6*. Hence, with regard to human mucin genes, the TR array of *MUC4* (as far as the longest alleles are concerned) is the

largest described so far. Eckhardt et al. [32] have recently reported that the TR domain of pig submaxillary mucin is encoded by an unusually long exon (34–32.8 kb).

The JER64 probe allowed us to isolate a *MUC4* cDNA clone from a human colon mucosa cDNA library. This clone contained a unique sequence located on the 5' side of the 48 bp TR domain. This sequence was extended by a 5' RACE procedure. Exon 1 consists of a 5' untranslated sequence and an 82 bp fragment encoding the first 27 N-terminal residues, which are very hydrophobic and might comprise the *MUC4* signal peptide as demonstrated with transcription and translation assays *in vitro* in the presence of dog pancreatic microsomes. An intron spanning at least 15 kb is located close to the boundary between the putative signal peptide and the mature peptide. Signal peptides are short N-terminal extension sequences required for the translocation of growing polypeptide chains through membranes of the endoplasmic reticulum [33]. In spite of the fact that these short cleaved N-terminal sequences constantly contain clustered hydrophobic residues flanked by hydrophilic residues, the primary sequences themselves vary considerably, in both length and structure [34–36]. A comparison of human and other mammalian preproteins therefore usually shows very little sequence similarity [37]. Nevertheless some proteins belonging to family groups show a high level of conservation through evolution of the signal peptide and occasionally of the 5' untranslated region sequence. For example, milk proteins (alpha S1, alpha S2 and beta caseins) [38], saliva proteins (proline-rich proteins) [39], statherin and histidine-rich peptides [40] are striking examples of conserved signal peptides. The high degrees of similarity (70–99%) of the signal peptide of the different rat and mouse proteins belonging to the proline-rich protein family is particularly striking, especially when considering that the tandem sequence repeats of these proteins highly diverge. Therefore, as the putative signal peptide of *MUC4* shows a high degree of similarity to that of ASGP-1, we can postulate that *MUC4* and rat ASGP-1 apomucins might have derived from a common ancestor. After the signal sequence, no similarities were found between the two proteins; however, the 3' cDNA sequence of *MUC4* remains to be studied to define whether other regions of *MUC4* demonstrate similarity to rat ASGP-1. However, for human statherin and basic histidine-rich peptide, two submandibular-gland proteins considered to be derived from a common ancestral gene, the majority of the coding sequence shows no significant similarity, except the first 14 N-terminal residues and the signal peptide. The unusual conservation of the sequence surrounding the AUG initiation codon has been proposed to contribute to specific regulation and/or to mRNA stability and conformation [39].

The second exon of *MUC4* encodes a peptide that can be divided into four distinct subdomains: (1) a first region with three imperfect repeats of 126 residues containing a high proportion of threonine and serine residues, (2) a second region of 554 residues corresponding to a unique sequence typically enriched for threonine and serine, (3) a third major domain, the 16-residue TR region, followed by (4) a fourth domain with a unique 26-residue sequence. In this regard *MUC4* resembles *MUC1* and *MUC2* genes. The N-terminal region of *MUC1* is also rich in serine, threonine and proline residues [24]. Similarly, the region upstream of the major TR domain of *MUC2* is rich in hydroxylated amino acid residues [30]. It consists mostly of 16-residue repeat units that are often non-continuous. As seen for *MUC4*, this region together with the TR array forms a single large exon.

Exon 2 also contains a tetrabasic amino acid RKRR site. The consensus sequence RXK/RR is conserved at the cleavage sites of a wide variety of secretory and membrane protein precursors

(e.g. von Willebrand factor, growth factors, serum proteases, receptors, and viral envelope glycoproteins) and is proposed to be processed by furin [41].

The four human mucin genes *MUC2*, *MUC5AC*, *MUC5B* and *MUC6* are clustered on chromosome 11p15.5 [42], whereas the five other genes are dispersed randomly throughout the human genome. A typical cysteine-rich domain, termed the Cys-subdomain, seems to be a characteristic of at least three (*MUC2*, *MUC5AC* and *MUC5B*) of the four human mucin genes located on 11p15.5 [17]. Desseyn et al. [43] suggested that these three genes might be a part of a highly conserved protein family and proposed an evolutionary tree.

No consensus cysteine-rich region was observed either in the N-terminus or in the TR array of *MUC4*, in contrast with the gel-forming mucins encoded by genes belonging to the 11p15.5 gene family [26,44–47]. Moreover, cell expression of *MUC4* distinguishes it from mucins located on 11p15.5. Thus *MUC4* is expressed in several epithelial cells other than typical mucin-secreting cells such as goblet and mucous cells; *MUC4* is expressed both in respiratory ciliated cells and in intestine absorptive cells [9].

Further investigations to elucidate the entire genomic organization of *MUC4* will be of great interest in shedding light on the dysregulation of this gene in tumorigenesis. The production of novel antibodies specifically defining *MUC4* apomucin is currently being performed and will greatly facilitate studies of *MUC4* protein/carbohydrate expression in epithelial disease.

We thank Annette Leclercq and Christine Mouton for performing polymorphism analysis, Marie-José Dejonghe for performing the SDS/PAGE analysis, Pascal Mathon for help in the preparation of the Figures, and Dr. Tor Savidge for his help in improving the style of this paper. This work was supported by le Comité du Nord de La Ligue contre le Cancer and l'Association de Recherche contre le Cancer. J. M. is a recipient of an Association Française de Lutte contre la Mucoviscidose fellowship.

REFERENCES

- Gendler, S. J. and Spicer, A. P. (1995) *Annu. Rev. Physiol.* **57**, 607–634
- Bansil, R., Stanley, E. and LaMont, J. T. (1995) *Annu. Rev. Physiol.* **57**, 635–657
- Forstner, G. (1995) *Annu. Rev. Physiol.* **57**, 585–605
- 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
- Crépin, M., Porchet, N., Aubert, J. P. and Degand, P. (1990) *Biorheology* **27**, 471–484
- Porchet, N., Nguyen, V. C., Dufossé, J., Audié, J. P., Guyonnet Dupérat, V., Gross, M. S., Denis, C., Degand, P., Bernheim, A. and Aubert, J. P. (1991) *Biochem. Biophys. Res. Commun.* **175**, 414–422
- Guyonnet Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Gallegue-Zouitina, S., Pigny, P., Degand, P., Aubert, J. P. and Porchet, N. (1995) *Biochem. J.* **305**, 211–219
- Gross, M. S., Guyonnet Dupérat, V., Porchet, N., Bernheim, A., Aubert, J. P. and Nguyen, V. C. (1992) *Ann. Genet.* **35**, 21–26
- Audié, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B. and Aubert, J. P. (1993) *J. Histochem. Cytochem.* **41**, 1479–1485
- Audié, J. P., Tetaert, D., Pigny, P., Buisine, M. P., Janin, A., Aubert, J. P., Porchet, N. and Boersma, A. (1995) *Hum. Reprod.* **10**, 98–102
- Balagué, C., Gambús, G., Carrato, C., Porchet, N., Aubert, J. P., Kim, Y. S. and Real, F. X. (1994) *Gastroenterology* **106**, 1054–1061
- Hollingsworth, M. A., Strawhecker, J. M., Caffrey, T. C. and Mack, D. R. (1994) *Int. J. Cancer* **57**, 198–203
- Walsh, M. D., McGuckin, M. A., Devine, P. L., Hohn, B. G. and Wright, R. G. (1993) *J. Clin. Pathol.* **46**, 922–925
- Nguyen, P. L., Niehans, G. A., Cherwitz, D. L., Kim, Y. S. and Ho, S. B. (1996) *Tumor Biol.* **17**, 176–192
- Ogata, S., Uehara, H., Chen, A. and Itzkowitz, S. H. (1992) *Cancer Res.* **52**, 5971–5978
- Wahl, G. M., Lewis, K. A., Ruiz, J. C., Rothenberg, B., Zhao, J. and Evans, G. A. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **4**, 2160–2164

- 17 Desseyn, J. L., Guyonnet Dupérat, V., Porchet, N., Aubert, J. P. and Laine, A. (1997) *J. Biol. Chem.* **272**, 3168–3178
- 18 Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- 19 Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472
- 20 Patthy, L. (1987) *FEBS Lett.* **214**, 1–7
- 21 Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148
- 22 Kyte, J. and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132
- 23 Wu, K., Fregien, N. and Carraway, K. L. (1994) *J. Biol. Chem.* **269**, 11950–11955
- 24 Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E. N. and Wilson, D. (1990) *J. Biol. Chem.* **265**, 15286–15293
- 25 Spicer, A. P., Parry, G., Patton, S. and Gendler, S. J. (1991) *J. Biol. Chem.* **266**, 15099–15109
- 26 Gum, Jr., J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1994) *J. Biol. Chem.* **269**, 2440–2446
- 27 Ohmori, H., Dohrman, A. F., Gallup, M., Tsuda, T., Kai, H., Gum, Jr., J. R., Kim, Y. S. and Basbaum, C. B. (1994) *J. Biol. Chem.* **269**, 17833–17840
- 28 Elhammer, A. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G. and Kezdy, F. J. (1993) *J. Biol. Chem.* **268**, 10029–10038
- 29 Carvalho, F., Seruca, R., David, L., Amorim, A., Seixas, M., Bennett, E., Clausen, H. and Sobrinho-Simoes, M. (1997) *Glycoconj. J.* **14**, 107–111
- 30 Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M. and Kim, Y. S. (1991) *J. Clin. Invest.* **88**, 1005–1013
- 31 Bobek, L. A., Tsai, H., Biesbrock, A. R. and Levine, M. J. (1993) *J. Biol. Chem.* **268**, 20563–20569
- 32 Eckhardt, A. E., Timpl, C. S., DeLuca, A. W. and Hill, R. L. (1997) *J. Biol. Chem.* **272**, 33204–33210
- 33 Kreil, G. (1981) *Annu. Rev. Biochem.* **50**, 317–348
- 34 Haeuptle, M. T., Flint, N., Gough, N. M. and Dobberstein, B. (1989) *J. Cell Biol.* **108**, 1227–1236
- 35 Bird, P., Gething, M. J. and Sambrook, J. (1990) *J. Biol. Chem.* **265**, 8420–8425
- 36 Sakaguchi, M., Tomiyoshi, R., Kuroiwa, T., Mihara, K. and Omura, T. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 16–19
- 37 Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. and Lingappa, V. R. (1979) *Symp. Soc. Exp. Biol.* **33**, 9–36
- 38 Gaye, P. and Mercier, J. C. (1981) *Reprod. Nutr. Dev.* **21**, 199–208
- 39 Clements, S., Mehansho, H. and Carlson, D. M. (1985) *J. Biol. Chem.* **260**, 13471–13477
- 40 Dickinson, D. P., Ridall, A. L. and Levine, M. J. (1987) *Biochem. Biophys. Res. Commun.* **149**, 784–790
- 41 Rehemtulla, A. and Kaufman, R. J. (1992) *Curr. Opin. Biotechnol.* **3**, 560–565
- 42 Pigny, P., Guyonnet Dupérat, V., Hill, A. S., Pratt, W. S., Galiegue-Zouitina, S., Collyn D'Hooge, M., Laine, A., Van Seuningen, I., Degand, P., Gum, J. R. et al. (1996) *Genomics* **38**, 340–352
- 43 Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P., Degand, P. and Laine, A. (1998) *J. Mol. Evol.* **46**, 102–106
- 44 Gum, Jr., J. R., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E. and Kim, Y. S. (1992) *J. Biol. Chem.* **267**, 21375–21383
- 45 Lesuffleur, T., Roche, F., Hill, A. S., Lacasa, M., Fox, M., Swallow, D. M., Zweibaum, A. and Real, F. X. (1995) *J. Biol. Chem.* **270**, 13665–13673
- 46 Desseyn, J. L., Aubert, J. P., Van Seuningen, I., Porchet, N. and Laine, A. (1997) *J. Biol. Chem.* **272**, 16873–16883
- 47 Toribara, N. W., Ho, S. B., Gum, E., Gum, Jr., J. R., Lau, P. and Kim, Y. S. (1997) *J. Biol. Chem.* **272**, 16398–16403