Human mucin gene *MUC5AC*: organization of its 5'-region and central repetitive region

Fabienne ESCANDE*†‡, Jean-Pierre AUBERT*†, Nicole PORCHET*†‡ and Marie-Pierre BUISINE*†¹

*INSERM Unité 377, place de Verdun, 59045 Lille Cedex, France, †Laboratoire de Biochimie et Biologie Moléculaire, Hôpital C. Huriez, Centre Hospitalier Régional et Universitaire, 59037 Lille Cedex, France, and ‡Faculté de médecine, Henri Warembourg place de Verdun, 59045 Lille Cedex, France

Human mucin gene *MUC5AC* is clustered with *MUC2*, *MUC5B* and *MUC6* on chromosome 11p15.5. We report here the full length cDNA sequence upstream of the repetitive region of human *MUC5AC*. We have also determined the sequence of its large central tandem repeat array. The 5'-region reveals high degree of sequence similarity with *MUC2* and *MUC5B* and codes for 1336 amino acids organized into a signal peptide, four pro-von Willebrand factor-like D domains (D1, D2, D' and D3) and a short domain which connects to the central repetitive region. In the central region, 17 major domains have been identified. Nine code for cysteine-rich domains (Cys-domains 1–9) and exhibit high sequence similarity to the cysteine-rich domains described in the central region of *MUC2* and *MUC5B*. Cys-domains 1–5 are interspersed by domains enriched with serine, threonine, and proline residues. Cys-domains 1–9 are

interspersed by four domains (TR1–TR4) composed of various numbers of *MUC5AC*-type repeats. Southern-blot analyses reveal allelic variations both in length and nucleotide sequence. The length polymorphism which is due to variable numbers of tandem repeats is located in TR1 and TR4, whereas a mutation polymorphism detected with *TaqI* is located in Cys-domain 6. In this study, the organization of *MUC5AC* has been entirely elucidated showing extensive similarity to the other chromosome 11p15 *MUC* genes, particularly *MUC5B*, and providing additional arguments for common evolution from a single ancestral gene.

Key words: chromosome 11p15, von Willebrand factor, tandem repeat.

INTRODUCTION

The human mucin gene *MUC5AC* is clustered on chromosome 11p15.5 with *MUC2*, *MUC5B* and *MUC6* [1,2]. This family of genes is thought to encode the secreted gel-forming mucins, which give mucus its visco-elastic properties. They share several structural features including a large central repetitive region flanked by non-repeat regions. The central region is constituted of tandem repeats which vary in length from 24 nt in *MUC5AC* [3] to 507 nt in *MUC6* [4] and encode threonine- and serine-rich sequence. The 5'- and 3'-flanking regions encode cysteine-rich domains which share similarities with the B, C, D and CK domains found in the pro-von Willebrand factor (pro-vWF) [3,5–16]. Sequence similarities between these four genes support the notion that they have evolved from a single common ancestral gene [17]. Like the other *MUC* genes, they have been found to show a high level of genetically determined polymorphism which is mainly due to variation in the number of copies of tandemly repeated sequence (variable number tandem repeats; VNTR) within the large central domain [18].

Qualitative and quantitative alterations in the expression of *MUC5AC* gene have been reported in various inflammatory or neoplastic situations. *MUC5AC* which is not expressed in normal intestine is aberrantly expressed in colorectal adenomas [19,20]. Furthermore, the expression level of *MUC5AC* is linked to the degree of cellular dysplasia. Likewise, *MUC5AC* mRNAs are detectable in pancreatic cancers and not in normal pancreas [21]. In chronic airway diseases, such as asthma, bronchitis and cystic

fibrosis, an increase in mucus production is associated with the induction of *MUC5AC* expression, suggesting that it may contribute to the pathogenesis of hypersecretory respiratory disorders [22]. A potential role of *MUC5AC* during epithelial renewal, and differentiation during foetal development has also been discussed [23,24].

To date, the C-terminal region [10,11,13] and a part of the N-terminal region of *MUC5AC* [2,14,15] have been well characterized, but much less information is available regarding its central region. Several cDNA clones (JER54, JER58, JER47, JUL32, etc.) have been isolated in our laboratory by screening of two human tracheal cDNA phage libraries [3,25,26]. JER54 and JER58 are exclusively constituted of tandem repeat sequences. JER47 encodes two cysteine-rich domains surrounding a repetitive domain. It has not been possible to assign any of these *MUC5AC* cDNAs to particular positions in the coding sequence, but they are assumed to occupy positions in the central part of the gene. Moreover, analysis of these clones strongly suggests an alternating structure of cysteine-rich domains and repetitive domains in the central region of *MUC5AC*.

In this study, we report the complete cDNA sequence of the 5'-region of the human *MUC5AC* gene and the genomic sequence of its large central repetitive region. We hypothesized that HGM1 was a part of the 5'-region of *MUC5AC* due to the 99% sequence identity between the *MUC5AC* cDNA clones JUL32 [3] and HGM1 (positions nt 1942–2282) [12], and the > 60% similarity between the 5'-region of HGM1 and the D3-type domain of *MUC2* [5]. Moreover, a cDNA clone called MUCY

Abbreviations used: poly(A)⁺, polyadenylated; RT, reverse transcriptase; TR, tandem repeat domain; TSP, threonine/serine/proline-rich domain; VNTR, variable number tandem repeats; vWF, von Willebrand factor.

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

The nucleotide sequences reported here have been submitted to EMBL Nucleotide Sequence Database under the accession numbers AJ298317, AJ298318, AJ298319 and AJ292079.

was produced in our laboratory by 5'-rapid amplification of cDNA ends ('RACE')-PCR using a primer located in the 5'-region of *MUC5B*. In spite of sequence similarities, comparative sequence analyses between *MUCY* and the genomic *MUC5B* clone BEN1 showed that *MUCY* could not belong to *MUC5B*. We hypothesized that *MUCY* was a part of the 5'-region of *MUC5AC*.

Based on these two hypotheses, reverse transcriptase (RT)-PCR experiments were performed with oligonucleotides, sequences of which were deduced from HGM1 and *MUCY*. Overlapping clones were obtained, yielding the overall sequence of the 5'-region of *MUC5AC*. Moreover, based on the hypothesis that just as other mucin genes the central repetitive region of *MUC5AC* was derived from one exon [27–31], the organization of the central region was characterized by PCR using genomic DNA as template. Amplifications were performed with oligonucleotides located in the 5'- and 3'-regions of *MUC5AC* or in JER47.

EXPERIMENTAL

Oligonucleotide primers

Oligonucleotide primers used in amplification and sequence experiments were synthesized by Eurogentec (Liège, Belgium) or by MWG-Biotech (Ebersberg, Germany).

RT-PCR

The first RT-PCR reaction was performed using TITAN™ one tube RT-PCR kit (Roche Diagnostics, Meylan, France) according to the manufacturer's instructions. The reaction was conducted on human tracheal polyadenylated RNA [poly(A)⁺ RNA] (0.1 µg) (ClonTech, Heidelberg, Germany) using sense primer NAU164 located in 5'-region of *MUCY* clone with antisense primer NAU131, the sequence of which was deduced from the HGM1 cDNA clone (EMBL DataBank accession number X81649) (positions nt 623–643) [12].

For other RT-PCR experiments, human tracheal poly(A)⁺ RNA (ClonTech) (0.4 µg) was reverse transcribed using the Advantage™ RT-for-PCR kit (ClonTech) and random primers according to the manufacturer's instructions. First strand cDNA was amplified by PCR with various primers, sequences of which were deduced from HGM1 [NAU49 (positions nt 2445–2465), NAU117 (positions nt 566–586), NAU149 (positions nt 1109–1129), NAU189 (positions nt 1159–1179)], JUL32 (EMBL DataBank accession number Z34280) [NAU50 (positions nt 17–37), NAU51 (positions nt 150–170)] [3], and from JER54 (EMBL DataBank accession number AJ292079) [NAU34 (positions nt 46–93)] [25]. Amplifications were performed in a PerkinElmer Thermal Cycler 9700 (Applied Biosystems, Courtaboeuf, France). PCR reactions were conducted in 50 µl reaction volumes containing 5 µl of 10× buffer (100 mM Tris/HCl, 15 mM MgCl₂, 500 mM KCl, pH 8.3), 4 µl of 10 mM dNTPs, 1–2 µl of first strand tracheal cDNA, 10 pmol of each primer and 2 units of *Taq* DNA polymerase (Roche). The mixture was denatured at 94 °C for 4 min, followed by 30 cycles at 94 °C for 45 s, 58–60 °C for 45 s, and 72 °C for 2 min. The final elongation step was extended for an additional 15 min. After electrophoresis, the amplified products were purified using Wizard™ PCR Preps DNA purification resin (Promega, Charbonnières, France).

DNA cloning

PCR fragments were cloned into pCR2.1 vector with the Original TA cloning™ kit (Invitrogen, Leek, The Netherlands) according

to the manufacturer's instructions. Some restriction fragments were subcloned into the pBluescript KS(+) vector (Stratagene, St Quentin en Yvelines, France).

Plasmid DNA purification

Plasmid DNA was purified using Qiaprep Spin Plasmid kit (Qiagen, Courtaboeuf, France) or Wizard™ minipreps DNA purification system (Promega).

Amplifications on genomic DNA

PCR reactions were performed on genomic DNA from a unique volunteer (designated T97051) in the aim to limit variations relative to interindividual VNTR polymorphism.

To characterize the organization of the central region, long expansion PCR was performed using Expand™ Long Template PCR System (Roche) with primers chosen in specific non-repetitive sequences to avoid cross-hybridization with *MUC2* and *MUC5B* and multiple annealing: NAU149 (sense) located in HGM1 (positions nt 1109–1129), NAU182 (antisense) located in JER47 (EMBL DataBank accession number Z34277) (positions nt 925–948) [3] and NAU241 (antisense) located just downstream of the first intron in the 3'-region of *MUC5AC* (EMBL DataBank accession number AJ001402) (positions nt 670–690) [13]. The PCR amplification mixture (50 µl) contained 200 ng of genomic DNA, 2.5 µl of sodium, 10 mM dNTPs, 15 pmol of each primer, 1.5 µl of 2.25 mM MgCl₂, 5 µl of 10× Expand™ long Template PCR buffer 3 and 2.5 units of enzyme mixture. Amplifications were performed in a PerkinElmer Thermal Cycler 9700. PCR parameters were 94 °C for 2 min, followed by 30 cycles at 94 °C for 10 s, annealing at 60 °C for 45 s, and elongation at 71 °C for 45 s, and 68 °C for 10 min. The 20 last cycles had their elongation time extended by 10 s for each new cycle, followed by a final elongation at 71 °C for 15 min. PCR products were subjected to electrophoresis on agarose gel and stained with ethidium bromide. The band of interest was cut out and purified using QIAquick Gel Extraction kit (Qiagen) according to the manufacturer's instructions.

Additional PCR experiments were conducted as described above using 200 ng of genomic DNA as template with various primers located in HGM1 [NAU49 (positions nt 2445–2465), NAU117 (positions nt 566–586), NAU189 (positions nt 1159–1179)], JUL32 [NAU50 (positions nt 17–37)], and JER47 [NAU48 (positions nt 124–140), NAU608 (positions nt 102–122), NAU609 (positions nt 981–1001)].

Sequencing of DNA fragments and data analyses

Clones were sequenced on both strands by the dideoxy-chain termination method using the sequence 2.0 kit (U. S. Biochemical) according to the manufacturer's instructions using [α^{35} -S]dATP (Amersham International, Saclay, France) and either universal primers or specific internal oligonucleotides. Sequences were also determined by automatic sequencing 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) using standard vector primers or with ABI PRISM model 377 XL automatic sequencer using the ABI PRISM dRhodamine terminator cycle ready reaction kit (Applied Biosystems). Some of the sequencing analyses were performed by direct sequencing of the PCR products with specific primers. Analysis of nucleic acid and protein sequence data was performed using PC/Gene Software (Intelligenetics).

Figure 1 Strategy for the determination of sequence of the 5'-region and the central region of *MUC5AC*

(A) Strategy for obtaining the complete N-terminal cDNA of *MUC5AC*. *MUC5AC* cDNA clones from which primers were chosen are indicated. Fragments were obtained by RT-PCR. They were cloned into pCR2.1 vector and sequenced. To facilitate sequencing, TH164-131 was subcloned into pKS vector after digestion by restriction endonuclease *Pst*I (P). (B) Strategy for the determination of the sequence of the *MUC5AC* central region. Schematic representation of *MUC5AC* fragments obtained by PCR. Most of them have been cloned and entirely sequenced (continuous lines). (C) Representation of the deduced amino acid sequence of the 5'- and central regions. □, signal peptide; ▨, D-domains; ▩, MUC11p15-type domain; ■, cysteine-rich domains (Cys); ▤, threonine/serine/proline-rich domains (TSP); ▦, tandem repeat domain (TR); □, unique sequence.

Southern-blot analysis

Southern blots were performed on human genomic DNA from volunteer T97051 and on PCR products obtained by Expand long PCR on genomic DNA from T97051. Genomic DNA or PCR products were treated with restriction endonucleases (*Acc*I, *Hinf*I, *Sac*I, *Sau*I and *Taq*I; Roche). Restriction endonucleases were selected from the sequences flanking the repetitive region of *MUC5AC* and from the sequences of JER47 and JER58 (EMBL DataBank accession number Z34280) [3]. *Sac*I was expected to cut closest to the repeat region, *Hinf*I recognized a site within the sequence of JER58 and *Acc*I cut in the first cysteine-rich domain of JER47. *Sau*I and *Taq*I were expected to recognize sites within cysteine-rich domains. Fragments were separated by electrophoresis in phosphate buffer through 1% agarose gel, transferred on to a nylon N⁺ membrane and hybridized with the ³²P-labelled probes JER58 and TH50-49 as described previously [26]. JER58 probe is a cDNA probe only composed of *MUC5AC*-specific

tandem repeat units [3]. TH50-49 probe is a RT-PCR product specific to cysteine-rich domains (see the Results section for additional information).

RESULTS

Cloning and sequence analysis of the 5'-region of the *MUC5AC* tracheal cDNA

Since *MUC5AC* is highly expressed in trachea and has been initially isolated from a human tracheobronchial cDNA library, RT-PCR experiments were performed with human tracheal RNA.

The first RT-PCR reaction was conducted using the sense primer NAU164 located in MUCY with the antisense primer NAU131 located in HGM1 (see the Experimental section) (Figure 1A). This yielded a 3.7-kb PCR fragment designated TH164-131, 640-bp of which overlapped with HGM1. To facilitate sequence

Figure 2 Analysis of the PCR products ZH149-182 and ZH149-241

(A) PCR fragments were separated by electrophoresis in 0.8% agarose gel in borate buffer. M, marker DNA for which the sizes are indicated. (B) Southern blot of the agarose gel presented in (A) hybridized with 32 P-labelled *MUC5AC*-specific probes TH50-49 and JER58.

analysis, TH164-131 was subcloned after digestion by the restriction endonuclease *Pst*I. Another RT-PCR experiment was conducted using specific primers of HGM1 (NAU117 and NAU189), yielding a 0.6-kb fragment designated TH117-189.

The combined sequences from TH164-131 and TH117-189 allowed us to establish the whole coding sequence of the 5'-region of *MUC5AC*. The nucleotide sequence designed TH164-189 was 4228 bp in length [EMBL Databank under accession number AJ298317 (positions nt 1–4228)]. At base +39 there is a ATG codon embedded in a Kozak consensus sequence, constituting a putative translation start site [32]. Comparative sequence analyses reveal that the 5'-region of the *MUC5AC* tracheal cDNA is nearly similar (99.4%) to clone MUC5AC-5'RP isolated by Li et al. [14], except for 14 nt substitutions and two nucleotide insertions and deletions introducing shifts in the reading frame. The sequence is almost identical (99.2%) with clones HGM1 (positions nt 1–1179) and HGM2 [15].

The deduced peptide encoded by the open reading frame of the cDNA consists of 1370 amino acids. The sequence contains a high amount of cysteine residues (8.9%) and seven typical potential N-glycosylation sites (Asn-X-Ser/Thr, where X is any amino acid except Pro). It contains a potential signal peptide, four cysteine-rich domains similar to the D-domains of the human pro-vWF (353 aa, 368 aa, 100 aa and 390 aa) [16] and a short domain (92 aa) corresponding to the MUC11p15-type domain described by Desseyn et al. [9] (Figure 1C). Positions for the cysteine residues in the D-domains correspond almost exactly to those described previously for the N-termini of human MUC2 and MUC5B [5,6,9] and several other amino acid residues are also well conserved. The sequence similarity with MUC2 and MUC5B is about 50% and 59% respectively.

Cloning of the central region of *MUC5AC*

Characterization of the organization of the central region has not been an easy task because of its highly repetitive nature and

large size. Screening of genomic cosmid libraries with several specific *MUC5AC* probes did not yield any *MUC5AC* clone. RT-PCR and PCR experiments and Southern-blot analyses provided a convenient alternative to determine the organization and the sequence of the central region of *MUC5AC*.

First, RT-PCRs were conducted using specific primers of HGM1 (NAU49 and NAU149), JUL32 (NAU50 and NAU51) and JER54 (NAU34). Four overlapping clones, called TH149-51 (1 kb), TH50-49 (0.5 kb), TH50-49(2) (1.6 kb) and TH50-34 (1.3 kb) were obtained allowing us to make the physical link between the 5'-region of *MUC5AC* and its specific repetitive region (see their location in Figure 1A). The compilation of nucleotide sequences of these four clones yielded a sequence of 3225 bp, corresponding to HGM1 sequence (positions nt 1109–2551) extended by 1782 bp of novel sequence consisting of the perfect repetition except for three point mutations of the 3'-end sequence of HGM1 (positions nt 1433–2551) followed by 24 *MUC5AC*-type repeats.

Due to the difficulties of amplifying large fragments in the repetitive region by RT-PCR, PCR experiments were performed on genomic DNA using primers NAU149, NAU241, and NAU182 (see the Experimental section). We obtained two products of 10.5 kb and 5.5 kb which were called ZH149-241 and ZH149-182 respectively (Figures 1B and 2A). To verify that these fragments consisted of *MUC5AC* sequences, they were further characterized by Southern-blot analysis with *MUC5AC* specific probes JER58 and TH50-49. The probes reacted with the two PCR products (Figure 2B) indicating that they belonged to *MUC5AC*. The fragment ZH149-241 allowed us to make the physical link between the 5'- and 3'-regions of *MUC5AC*. The ZH149-241 and ZH149-182 fragments could not be successfully cloned into plasmid vectors. Nevertheless, these clones were partially sequenced from both ends by 'walking' with specific primers directly on PCR products. An additional PCR was performed using primers NAU117 and NAU189 (Figure 1B), and comparison of the genomic sequences determined with the cDNA

Figure 3 Amino acid sequence comparison between cysteine-rich domains of *MUC5AC*

Sequences were deduced from nucleotide sequences and were aligned to give maximum identity with a minimal number of gaps. —, gaps introduced in the sequence for alignment purposes. Cysteine residues are in **bold**. Highly conserved amino acids are shadowed and boxed.

sequences revealed the presence of the last intron upstream the tandem repeat array, thus identifying the 5'-end of the central exon. It is 687 bp in length (results not shown) and located in the *MUC11p15*-type domain between positions nt 4047 and nt 4048 (class I). With our previous results regarding the 3'-region of *MUC5AC* [13], this result allowed the determination of the boundaries of the large central exon.

Next, Southern-blot analysis of the two PCR products ZH149-241 and ZH149-182 was performed after digestion by restriction endonucleases *HinfI*, *SacI* and *AccI* (results not shown). Comparison of profiles obtained after hybridization with JER58 and ZH50-49 probes allowed us to ascertain the organization of the repetitive region as follows: the size of the large central exon is approx. 10.5 kb, *HinfI* which only cuts inside the ZH149-182 fragment permitted us to position JER58 in this region, and likewise, *AccI* which cuts only once inside the ZH149-241 fragment permitted us to precisely localize the JER47 clone.

Further study involved PCR experiments using primers located in HGM1 (NAU49), JUL32 (NAU50) and JER47 (NAU48, NAU608 and NAU609). Four overlapping fragments were obtained and called ZH50-608 (3.8 kb), ZH48-49 (1.1 kb), ZH609-49 (1.5 kb), and ZH48-49(2) (2.3 kb) respectively (Figure 1B). Clones ZH48-49 and ZH609-49 were fully sequenced, but only end-sequences of ZH50-608 and ZH48-49(2) could be obtained. Nevertheless, these regions likely consist entirely of tandem repeats as strongly suggested by the results of Southern-blot analysis performed on genomic DNA from T97051 (see below and Figure 5). The four clones show a particular internal organization similar to that of JER47. Two non-repetitive sequences are located at the 5'-end and at the 3'-end of the clone, surrounding a 24 bp tandem repeat sequence which characterizes the *MUC5AC* gene. ZH48-49 shows 100% sequence similarity to JER47 [3] and ZH50-608 contains the sequence of clone 4F described previously by Ho et al. [33].

Additional PCR experiments were carried out using primers NAU917 and NAU918 chosen in the repetitive sequence of ZH609-49, allowing us to obtain three overlapping fragments [ZH48-918 (1.3 kb), ZH917-49 (1.4 kb) and ZH917-241 (3 kb)] (Figure 1B) which were entirely sequenced and permitted us to confirm our results.

Sequence analyses of the central region of *MUC5AC* and of the deduced peptide

The compiled nucleotide sequences obtained allowed us to establish the near complete sequence of the central region. The nucleotide sequences are available from EMBL DataBank under accession numbers AJ298317 (positions nt 4229–7382), AJ298318 and AJ298319.

The size of the central exon is approx. 10.5 kb. In the deduced amino acid sequence, 17 major domains have been identified (Figure 1C). Nine code for cysteine-rich domains, called Cys-domains 1–9, which contain ten cysteine residues per 110 amino acid residues. Their amino acid sequences are displayed in Figure 3. In addition to the remarkable conservation of the cysteine residues, the conservation of numerous other amino acid residues is to be noted. Cys-domain 2 and Cys-domain 4 present a perfect sequence similarity except for five mutations at the nucleotide level which lead to only two modifications in the deduced amino acid sequence. Cys-domain 3 and Cys-domains 5–9 are nearly identical and contain the TR3A tryptic peptide (WFDVD-FPSPGPHGGDKETYNNI) which was initially isolated from human lung mucus [34]. There is one conserved potential O-glycosylation site in each of these subdomains. No potential N-glycosylation site exists. Cys-domains 1–5 are interspersed with no repetitive sequences, but threonine/serine/proline-rich domains designated TSP1–TSP4 which contain no cysteine residues. TSP1/TSP3 and TSP2/TSP4 are nearly identical and only include three and two modifications in the deduced amino acid sequence respectively. Five domains Cys-domains 5–9 are interspersed by four domains (TR1–TR4) composed of various numbers of *MUC5AC*-specific tandem repeat of 8 aa. The alignment of the corresponding deduced amino acid sequences of TR1–TR4 is shown in Figure 4. Domains TR1 and TR4 are comprised of approx. 124 tandem repeats and 66 tandem repeats respectively, whereas TR2 is composed of 17 tandem repeats, and TR3 of 34 tandem repeats. Although the consensus sequence TTSTTSAP is found more frequently, other motifs such as GTTPSPVP are also frequently observed. It is interesting to note that one of the tandem repeat units within TR3 is only 7 aa. This result did not correspond to an artifact since it was also obtained from genomic DNA from another individual. Many serine/

Figure 5 Southern-blot analysis of genomic DNA from T97051 individual

(A) Restriction enzyme digested genomic DNA was hybridized with 32 P-labelled probes JER58 and TH50-49. (B) Southern-blot results and sizes of digested fragments hybridized with JER58 or TH50-49 and corresponding alleles (A1, Allele 1; A2, Allele 2). (C) Partial restriction map of the central region of *MUC5AC* as determined by Southern-blot analysis. The two alleles are represented. A, *AccI*; H, *HinfI*; SI, *SacI*; S, *SauI*; T, *TaqI*. The locations of some well known *MUC5AC* cDNA clones are also indicated.

confirmed by digestion with *TaqI* (A1, 4.1 kb; A2, 3.5 kb) and *SauI* (A1, 3.3 kb; A2, 3.45 kb), which cut within Cys-domain 5 and Cys-domain 6 located at the both ends of the first

tandem repeat domain and also by double digestions *HinfI/TaqI* (A1, 2.9 kb; A2, 1.95 kb) and *HinfI/SauI* (A1, 2 kb; A2, 2.1 kb). In the same way, comparison of the two profiles obtained after

digestion by *HinfI*/*TaqI* or *HinfI*/*SauI* and hybridization with JER58 strongly suggested the presence of a restriction site polymorphism *TaqI* in the Cys-domain 6 (*HinfI*/*TaqI*: A1, 2.9 kb; A2, 1.95 kb) (*HinfI*/*SauI*: A1, 2.1 kb; A2, 2 kb). Sequencing of the ZH48-49 clones corresponding to the two alleles (see their location in Figure 1B) confirmed this result, and allowed us to conclude that TR2 did not vary in length in our individual.

Comparison of the profiles obtained after digestion by *TaqI* and *SauI* permitted us to confirm the results obtained previously. They allowed us also to establish that TR3 did not vary in length (*SauI*, 1.25 kb; *TaqI*, 1.25 kb for A1 and A2), whereas TR4 presented a VNTR polymorphism (*SauI*: A1, 2.05 kb; A2, 2.1 kb) and (*TaqI*: A1, 1.25 kb; A2, 1.95 kb) and that TR1 and TR4 domain consisted entirely of repeats. The precise partial restriction map of the central region of *MUC5AC* that could be deduced from results from Southern-blot analyses and also from nucleotide sequence is indicated in Figure 5C.

DISCUSSION

The study reported here presents the overall sequence of the 5'- and central region of *MUC5AC*. The sequence of the 5'-region of the *MUC5AC* tracheal cDNA was performed by RT-PCR. Comparative sequence analysis between *MUC5AC*, *MUC2*, *MUC5B* and *vWF* revealed strikingly common features which permitted us to suppose that at base +39 was the translation initiation site. This hypothesis was confirmed by Li et al. [14]. The N-terminal region of *MUC5AC* contains, as in *MUC2* and *MUC5B*, four cysteine-rich domains which share significant amino acid identities with the D-domains present in the N-terminal region of the pro-vWF (Figure 6). Nearly all the cysteine residues and numerous amino acids are conserved. Multimerization of pro-vWF requires D-domains found in the N-terminal region and particularly three CGLC motifs located in D1-, D2- and D3-domains which are thought to play an important role in the catalysis of interchain disulphide bond formation [36]. As these three motifs are perfectly conserved in D-domains of *MUC5AC*, it is likely that D-domains found in the N-terminal region of *MUC5AC* may also function in disulphide-linked multimer formation as shown for porcine submaxillary mucin [37].

MUC genes are characterized by a large central region which contains tandem repeats rich in serine and threonine residues supporting numerous O-glycan chains and which are encoded by a single large exon [27–31]. In *MUC5AC*, the size of the central exon is evaluated to be approx. 10.5 kb in our individual. The determination of its boundaries permit us to identify the position of the last intron (687 bp) of the 5'-region of *MUC5AC* which is identical to that of the last intron in the 5'-region of *MUC5B* [9]. Thus, although we did not determine the entire genomic

organization of the 5'-region of *MUC5AC*, it is likely identical to that of *MUC5B* [9] as we demonstrated previously for the 3'-region of *MUC5AC* [13].

The organization of the central exon has been determined by Southern-blot analysis and PCR experiments. As it has been encountered with other repetitive mucin clones, some of the clones were unstable when attempts were made to clone them into plasmid vectors and could not be entirely sequenced [28]. This appears to be due to the unstable nature of DNA segments containing tandem repeats in bacterial host strains. Nevertheless, the precise organization of the central region could be determined, and 17 major domains have been identified in the deduced peptide (Figure 6). Nine domains (Cys-domains 1–9) of 110 amino acid residues contain ten cysteine residues, and show a very similar organization to that observed twice in *MUC2* and rat *Muc2*, at least twice in mouse *Muc5ac* and seven times in *MUC5B* [17]. The structure of the cysteine-rich domains has been conserved over a long evolutionary time scale. The conservation of numerous amino acid residues, in addition to the cysteine residues, suggests an important role for these domains in the preservation of the essential function of mucin. Four repetitive domains (TR1–TR4) are comprised of *MUC5AC*-specific repeats of 8 aa which, unlike cysteine-rich domains, differ in sequence and size between both species and different mucins. Most of the mucin genes tested have been found to show a high level of genetically determined polymorphism [31,38]. This polymorphism is mainly due to VNTR within the large central domain and is thus reflected both at the RNA and DNA levels [18,38–40]. In some cases, the variation in repeat numbers is so great that it could be responsible for a 2-fold difference in the length of the coding sequence and size of the protein [38]. For example, *MUC1* contains from as few as 20 to as many as 125 tandem repeats in the population [41], the *MUC4* gene from approx. 145–395 repeats [31] and the *MUC2* gene from approx. 51–115 repeats [28]. These VNTR polymorphisms are thought to occur by unequal, but homologous, cross-over events. In contrast, no VNTR has been demonstrated in the central region of *MUC5B*. Restriction-fragment-length polymorphism of the *MUC5AC* gene has been noted in our laboratory and by others [38,42]. Knowledge of the *MUC5AC* gene structure obtained in this study will allow us to better understand the individual differences that exist in this gene. We have demonstrated that the central region is coded by approx. 10.5 kb in our individual and that a minimal degree of VNTR genetic polymorphism (in our individual, the two alleles differs in size by less than 0.5 kb) is observed, as described previously [38]. The small difference in length of *MUC5AC* and no length variation of *MUC5B* could be explained because *MUC5AC* and *MUC5B* have interrupting cysteine-rich domains, seven in the case of *MUC5B* and nine in *MUC5AC*, whereas the other mucins contain a major uninterrupted tandem repeat. It seems that the

Figure 6 Representation of the entire peptide structure of *MUC5AC*

□, signal peptide; ▨, D-domains; ▩, MUC11p15-type domain; ■, cysteine-rich domains (Cys); ▤, threonine/serine/proline-rich domains (TSP); ▧, tandem repeat domain (TR); □, unique sequence; ▨, A3uD4-domain; ▩, B-domain; ▤, C-domain; ▧, cysteine-knot domain.

first type of organization limits the VNTR mutation and stabilizes the gene structure, whereas the long stretches of simple repeats are conducive to VNTR mutation because of homologous recombination. According to this, in *MUC5AC*, the two polymorphic zones are located in TR1 and TR4 which contain more tandem repeat units than TR2 or TR3 and also more tandem repeat units than in *MUC5B*. The *MUC5AC* gene also exhibits a mutation polymorphism in the Cys-domain 6 that is detectable using the enzyme *TaqI*. A similar polymorphism for *TaqI* has been characterized for *MUC6* and *MUC2* [29,39].

Variation in length of the glycosylated tandem repeat domain is likely to affect mucin functions by altering the physicochemical properties through changes in the size and extent of glycosylation, and thus the accessibility of the cell surface to bacteria and other environmental agents, as well as the consistency of the mucus. For instance it has been demonstrated for *MUC1* and rat *Muc4* that the number of repeats is a critical factor for their anti-adhesive property [43,44]. Nevertheless, it is not clear at this stage what impact variation in the length of the tandem repeat region of *MUC5AC* is likely to have, especially since the predicted differences in polypeptide length are not substantial.

In conclusion, the *MUC5AC* cDNA sequence is approximately 17.5 kb in length with a central exon of 10.5 kb containing the overall repetitive region. Its organization is similar to those of the other members of the chromosome 11p15 mucin gene family, particularly *MUC5B*, providing additional arguments for their common evolution from a single ancestral gene.

This work was supported by the Association de Recherche contre le Cancer and by the EC Concerted Action (grant number BMH4-CT98-3222). We thank A. Laine and J.-L. Desseyn for providing the cDNA clone MUCY. We thank A. Leclercq, D. Petitprez, K. Rousseau, P. Mathon, and V. Mortelee for excellent technical assistance.

REFERENCES

- Lagow, E., DeSouza, M. M. and Carson, D. D. (1999) Mammalian reproductive tract mucins. *Hum. Reprod.* **5**, 280–292
- Pigny, P., Guyonnet-Dupérat, V., Hill, A. S., Pratt, W. S., Galiègue-Zouitina, S., D'Hooge, M. C., Laine, A., Van-Seuning, I., Degand, P., Gum, J. R. et al. (1996) Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. *Genomics* **38**, 340–352
- Guyonnet-Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Galiègue-Zouitina, S., Pigny, P., Degand, P., Aubert, J. P. and Porchet, N. (1995) Characterization of the human mucin gene *MUC5AC*: a consensus cysteine-rich domain for 11p15 mucin genes? *Biochem. J.* **305**, 211–219
- Toribara, N. W., Robertson, A. M., Ho, S. B., Kuo, W. L., Gum, E., Hicks, J. W., Gum, Jr, J. R., Byrd, J. C., Siddiki, B. and Kim, Y. S. (1993) Human gastric mucin. Identification of a unique sequence by expression cloning. *J. Biol. Chem.* **268**, 5879–5885
- Gum, Jr, J. R., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E. and Kim, Y. S. (1992) The human *MUC2* intestinal mucin gene has cysteine-rich subdomains located upstream and downstream of its central repetitive region. *J. Biol. Chem.* **267**, 21375–21383
- Gum, Jr, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1994) Molecular cloning of human intestinal mucin (*MUC2*) cDNA. Identification of the amino terminus and overall sequence similarity to prepro-Willebrand factor. *J. Biol. Chem.* **269**, 2440–2446
- Toribara, N. W., Ho, S. B., Gum, E., Gum, Jr, J. R., Lau, P. and Kim, Y. S. (1997) The carboxy-terminal sequence of the human secretory mucin, *MUC6*. Analysis of the primary amino acid sequence. *J. Biol. Chem.* **272**, 16398–16403
- Desseyn, J. L., Aubert, J. P., Van Seuning, I., Porchet, N. and Laine, A. (1997) Genomic organization of the 3' region of the human mucin gene *MUC5B*. *J. Biol. Chem.* **272**, 16873–16883
- Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P. and Laine, A. (1998) Genomic organization of the human mucin gene *MUC5B*. cDNA and genomic sequences upstream of the large central exon. *J. Biol. Chem.* **273**, 30157–30164
- Merzaman, D., Charles, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M. and Rose, M. C. (1994) Cloning and analysis of cDNA encoding a major airway glycoprotein human tracheobronchial mucin (*MUC5*). *J. Biol. Chem.* **269**, 12932–12939
- Lesuffleur, T., Roche, F., Hill, A. S., Lacara, M., Fox, M., Swallow, D. M., Zweibaum, A. and Real, F. X. (1995) Characterization of a mucin cDNA clone isolated from HT-29 mucus secreting cells. The 3' end of *MUC5AC*? *J. Biol. Chem.* **270**, 13665–13673
- Klomp, L. W., Van Rens, L. and Strous, G. J. (1995) Cloning and analysis of human gastric mucin cDNA reveals two types of conserved cysteine-rich domains. *Biochem. J.* **308**, 831–838
- Buisine, M. P., Desseyn, J. L., Porchet, N., Degand, P., Laine, A. and Aubert, J. P. (1998) Genomic organization of the 3' region of the human *MUC5AC* mucin gene: additional evidence for a common ancestral gene for 11p15.5 mucin gene family. *Biochem. J.* **332**, 729–738
- Li, D., Gallup, M., Fan, N., Szymkowski, D. E. and Basbaum, C. B. (1998) Cloning of the amino-terminal and 5'-flanking region of the human *MUC5AC* mucin gene and transcriptional up-regulation by bacterial exoproducts. *J. Biol. Chem.* **273**, 6812–6820
- van den Bovenkamp, J. H., Hau, C. M., Strous, G. J., Buller, H. A., Dekker, J. and Einerhand, A. W. (1998) Molecular cloning of human gastric mucin *MUC5AC* reveals conserved cysteine-rich D-domains and a putative leucine zipper motif. *Biochem. Biophys. Res. Commun.* **245**, 853–859
- Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. and Sadler, J. E. (1989) Structure of the gene for human von Willebrand factor. *J. Biol. Chem.* **264**, 19514–19527
- Desseyn, J. L., Aubert, J. P., Porchet, N. and Laine, A. (2000) Evolution of the large secreted gel-forming mucins. *Mol. Biol. Evol.* **17**, 1175–1184
- Debailleul, V., Laine, A., Huet, G., Mathon, P., D'Hooghe, M. C., Aubert, J. P. and Porchet, N. (1998) Human mucin gene *MUC2*, *MUC3*, *MUC4*, *MUC5AC*, *MUC5B*, and *MUC6* express stable and extremely large mRNAs and exhibit a variable length polymorphism. An improved method to analyse large mRNAs. *J. Biol. Chem.* **273**, 881–890
- Buisine, M. P., Janin, A., Maunoury, V., Audié, J. P., Delescaut, M. P., Copin, M. C., Colombel, J. F., Degand, P., Aubert, J. P. and Porchet, N. (1996) Aberrant expression of a human mucin gene (*MUC5AC*) in rectosigmoid villous adenoma. *Gastroenterology* **110**, 84–91
- Bartman, A. E., Sanderson, S. J., Ewing, S. L., Niehans, G. A., Wiehr, C. L., Evans, M. K. and Ho, S. B. (1999) Aberrant expression of *MUC5AC* and *MUC6* gastric mucin genes in colorectal polyps. *Int. J. Cancer* **80**, 210–218
- Balague, C., Audié, J. P., Porchet, N. and Real, F. X. (1995) *In situ* hybridization shows distinct patterns of mucin gene expression in normal, benign, and malignant pancreas tissues. *Gastroenterology* **109**, 953–964
- Temann, U. A., Prasad, B., Gallup, M. W., Basbaum, C., Ho, S. B., Flavell, R. A. and Rankin, J. A. (1997) A novel role for murine IL-4 *in vivo*: induction of *MUC5AC* gene expression and mucin hypersecretion. *Am. J. Respir. Cell Mol. Biol.* **16**, 471–478
- Buisine, M. P., Devisme, L., Savidge, T. C., Gespach, C., Gosselin, B., Porchet, N. and Aubert, J. P. (1998) Mucin gene expression in human embryonic and foetal intestine. *Gut* **43**, 519–524
- Buisine, M. P., Devisme, L., Copin, M. C., Durand-Réville, M., Gosselin, B., Aubert, J. P. and Porchet, N. (1999) Developmental mucin gene expression in the respiratory tract. *Am. J. Respir. Cell Mol. Biol.* **20**, 209–218
- Aubert, J. P., Porchet, N., Crepin, M., Dutergue-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D. and Degand, P. (1991) Evidence for different human tracheobronchial mucin peptides deduced from nucleotide cDNA sequences. *Am. J. Respir. Cell Mol. Biol.* **5**, 178–185
- 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) Molecular cloning and chromosomal localization of a novel tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem. Biophys. Res. Commun.* **175**, 414–422
- Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitou, J. and Gendler, S. (1990) Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochem. Biophys. Res. Commun.* **173**, 1019–1029
- Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M. and Kim, Y. S. (1991) *MUC-2* human small intestinal mucin gene structure. Repeated arrays and polymorphism. *J. Clin. Invest.* **88**, 1005–1013
- Bobek, L. A., Liu, J., Sait, S. N. J., Shows, T. B., Bobek, Y. A. and Levine, M. J. (1996) Structural and chromosomal localization of the human salivary mucin gene *MUC7*. *Genomics* **31**, 277–282
- Desseyn, J. L., Guyonnet-Dupérat, V., Porchet, N., Aubert, J. P. and Laine, A. (1997) Human mucin gene *MUC5B*: the 10.7 kb large central exon encodes various subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. *J. Biol. Chem.* **272**, 3168–3178
- Nollet, S., Moniaux, N., Maury, J., Petitprez, D., Degand, P., Laine, A., Porchet, N. and Aubert, J. P. (1998) Human mucin gene *MUC4*: organization of its 5'-region and polymorphism of its central tandem repeat array. *Biochem. J.* **332**, 739–748
- Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148

- 33 Ho, S., Robertson, A., Shekels, L., Lyftogt, C., Niehans, G. and Toribara, N. (1995) Expression cloning of gastric mucin complementary DNA and localization of mucin gene expression. *Gastroenterology* **109**, 735–747
- 34 Rose, M. C., Kaufman, B. and Martin, B. M. (1989) Proteolytic fragmentation and peptidic mapping of human carboxyamidomethylated tracheobronchial mucin. *J. Biol. Chem.* **264**, 8193–8199
- 35 Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L. and Brunak, S. (1998) NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconjugate J.* **15**, 115–30
- 36 Mayadas, T. N. and Wagner, D. D. (1992) Vicinal cysteines in the prosequence play a role in von Willebrand factor multimer assembly. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 3531–3535
- 37 Perez-Vilar, J. and Hill, R. L. (1998) Identification of the half-cystine residues in porcine submaxillary mucin critical for multimerization through the D-domains. Roles of the GGLCG motif in D1- and D3-domains. *J. Biol. Chem.* **273**, 34527–34534
- 38 Vinal, L. E., Hill, A. S., Pigny, P., Pratt, W. S., Toribara, N., Gum, J. R., Kim, Y. S., Porchet, N., Aubert, J. P. and Swallow, D. M. (1998) Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Hum. Genet.* **102**, 357–366
- 39 Lan, M. S., Batra, S. K., Qi, W. N., Metzgar, R. S. and Hollingsworth, M. A. (1990) Cloning and sequencing of a human pancreatic tumor mucin cDNA. *J. Biol. Chem.* **265**, 15294–15299
- 40 Gross, M. S., Guyonnet-Dupérat, V., Porchet, N., Bernheim, A., Aubert, J. P. and Nguyen, V. C. (1992) Mucin 4 (MUC4) gene: regional assignment (3q29) and RFLP analysis. *Ann. Genet.* **35**, 21–26
- 41 Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J., Duhig, J., Peat, T., Burchell, J. M., Pemberton, L., Lalani, E. N. and Wilson, D. (1990) Molecular cloning and expression of a human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* **265**, 15286–15293
- 42 Pigny, P., Pratt, W. S., Laine, A., Leclercq, A., Swallow, D. M., Nguyen, V. C., Aubert, J. P. and Porchet, N. (1995) The MUC5AC gene: RFLP analysis with the JER58 probe. *Hum. Genet.* **96**, 367–368
- 43 Zrihan-Licht, S., Vos, H. L., Baruch, A., Elroy-Stein, O., Sagiv, D., Keydar, I., Hilken, J. and Wreschner, D. H. (1994) Characterization and molecular cloning of a novel MUC1 protein, devoid of tandem repeats, expressed in human breast cancer tissue. *Eur. J. Biochem.* **224**, 787–795
- 44 Komatsu, M., Yee, L. and Carraway, K. L. (1999) Overexpression of sialomucin complex, a rat homologue of MUC4, inhibits tumor killing by lymphokine-associated killer cells. *Cancer Res.* **48**, 2214–2220

Received 14 February 2001/29 May 2001; accepted 10 July 2001