



**HAL**  
open science

## Sequence of the 5'-flanking region and promoter activity of the human mucin gene MUC5B in different phenotypes of colon cancer cells

Isabelle van Seuningen, Michaël Perrais, Pascal Pigny, Nicole Porchet,  
Jean-Pierre Aubert

### ► To cite this version:

Isabelle van Seuningen, Michaël Perrais, Pascal Pigny, Nicole Porchet, Jean-Pierre Aubert. Sequence of the 5'-flanking region and promoter activity of the human mucin gene MUC5B in different phenotypes of colon cancer cells. *Biochemical Journal*, 2000, 348, pp.675 - 686. hal-02905708

**HAL Id: hal-02905708**

**<https://hal.science/hal-02905708v1>**

Submitted on 29 Sep 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Sequence of the 5'-flanking region and promoter activity of the human mucin gene  
MUC5B in different phenotypes of colon cancer cells.**

Isabelle Van Seuningen<sup>1</sup>, Michaël Perrais<sup>1</sup>, Pascal Pigny<sup>1,3</sup>, Nicole Porchet<sup>1,2</sup> and Jean-Pierre  
Aubert<sup>1,2</sup>

<sup>1</sup>Unité INSERM 377, Place de Verdun, 59045 Lille Cedex, France, Laboratoires de  
<sup>2</sup>Biochimie et de Biologie Moléculaire et de <sup>3</sup>Biochimie Endocrinienne de l'Hôpital C.  
Huriez, CHRU, 59037 Lille Cedex, France.

Short title : MUC5B gene expression in colon cancer cells

Section : Gene structure and expression

Key words : MUC5B, colon, cancer, promoter, mucins, differentiation

To whom correspondence should be sent :

Isabelle Van Seuningen

Unité INSERM 377, Place de verdun, 59045 Lille Cedex, France

Phone : 33. 320.29.88.65

FAX : 33.320.53.85.62

E-mail : [vanseuni@lille.inserm.fr](mailto:vanseuni@lille.inserm.fr)

Sequences submitted to GenBank : GenBank Accession Numbers AJ011582 and AJ012453

## **Abbreviations**

aa : amino acid

DMEM : Dulbecco's modified Eagle's minimum essential medium

GAPDH : glyceraldehyde-3-phosphate dehydrogenase

MTX : methotrexate

RT : reverse transcription

STD : standard

## Synopsis

Control of gene expression in intestinal cells is poorly understood. Molecular mechanisms that regulate transcription of cellular genes are the foundation for understanding developmental and differentiation events. Mucin gene expression has been shown to be altered in many intestinal diseases and especially cancers of the gastrointestinal tract. Toward understanding the transcriptional regulation of a member of the 11p15.5 human mucin gene cluster, we have characterised 3.55 kb of the 5'-flanking region of the human mucin gene *MUC5B* that include the promoter, the first two exons and the first intron. We report here the promoter activity of successively 5'-truncated sections of 956 bases of this region by fusing it to the coding region of a luciferase reporter gene. The transcription start site was determined by primer extension analysis. The region upstream of the transcription start site is characterised by the presence of a TATA box at bases -32/-26, DNA binding elements for transcription factors c-Myc, N-Myc, Sp1 and NF- $\kappa$ B as well as putative AP-1, CREB, HNF-1, HNF-3, TGT3, GKLF, TTF-1 and GRE binding sites. Intron 1 of *MUC5B* was also characterised, it is long of 2511 nucleotides and contains a DNA segment of 259 base pairs in which are clustered 8 tandemly repeated GA boxes and a CACCC box that bind Sp1. AP-2 $\alpha$  and GATA-1 nuclear factors were also shown to bind to their respective cognate element in intron 1. In transfection studies *MUC5B* promoter showed a cell-specific activity as it is very active in mucus-secreting LS174T cells whereas it is inactive in Caco-2 enterocytes and HT-29 STD undifferentiated cells. Within the promoter, maximal transcription activity was found in a segment covering the first 223 base pairs upstream of the transcription start site. Finally, in co-transfection experiments a transactivating effect of Sp1 onto *MUC5B* promoter was seen in LS174T and Caco-2 cells.

## Introduction

Human mucins are the main glycoprotein component of epithelial secretions and play important roles in maintaining mucosae integrity. So far, 8 genes have been described [1] and among them *MUC6-MUC2-MUC5AC-MUC5B* are clustered on human 11p15.5 chromosome [2]. Previous studies have demonstrated that the expression of these four genes is tissue- and cell-specific, and that their expression is dysregulated during the pathogenesis of several diseases [3]. These results suggest that mucin gene expression is tightly regulated. *MUC5B* is mainly expressed in normal bronchus, submaxillary glands, endocervix, pancreas [4], gall bladder [5,6]. Interestingly enough, *MUC5B* has been shown to be highly expressed in colon cancer [3] and especially in the mucus-secreting colon cancer cell lines HT-29 MTX (methotrexate) [7] and LS174T [6].

The *MUC5B* gene is 39.09 kb in length, contains 49 exons and encodes a 5701 amino acid (aa) polypeptide with an  $M_r$  of approximately 627,000. *MUC5B* shows extensive similarity in 5' and 3' regions to the other genes of the cluster *MUC2*, *MUC5AC* and *MUC6* [8-11]. Regarding *MUC5B* regulation, we previously showed that the intron G, now called intron 37 (30 kb downstream of the ATG), contains 8 repeats of a perfect Sp1 (GGGCGG) binding site [9]. It binds Sp1 and a 42 kDa nuclear factor called NF1-*MUC5B* specific to the mucus-secreting colon cancer cell line HT-29 MTX [12].

Regarding the other 11p15.5 mucin genes, promoter sequences and transcriptional regulation are known for *MUC2* and *MUC5AC*. *MUC2* a prominent mucin gene expressed in the colon, has a TATA and a CCAAT box as well as predicted binding sites for many transcription factors [13,14]. Transfection studies carried out in colon cancer cell lines (HT-29, LS174T) showed that a region in the promoter of *MUC2* that extends 848 bp upstream of the ATG confers maximal transcriptional activity [13]. Gum *et al.* showed that a

short segment including bases -91/-73 relative to the start site was essential for basal promoter activity in several colon cancer cell lines [14]. By the means of transgenesis, the same group showed that *cis*-elements necessary for *MUC2* goblet cell-specific expression are present within *MUC2* promoter [15]. More recently, it was shown that Sp1 regulates rat *muc2* gene transcription in lung cells [16]. Li *et al.* showed that *Pseudomonas aeruginosa* lipopolysaccharide activates a c-Src-Ras-MEK 1/2-pp90rsk signalling pathway that leads to activation of nuclear factor NF- $\kappa$ B (p65/p50) [17]. Activated NF- $\kappa$ B binds to  $\kappa$ B sites in the 5' proximal region of the gene and induces *MUC2* transcription [18,19].

*MUC5AC* is highly expressed in airway epithelial cells and in the stomach. A TATA box is located at -23/-29 of the transcription start site and putative binding sites for Sp1, AP-2, GRE, NF- $\kappa$ B as well as a CACCC box were found within the promoter [20]. *MUC5AC* transcription is activated when bronchial explants, airway epithelial cells (NCIH292) or colon cancer cells (HM3) are exposed to *Pseudomonas aeruginosa* or its exoproducts in cell-free supernatants. Both *MUC2* and *MUC5AC* are up-regulated by other Gram-positive and Gram-negative organisms [20]. These results suggest that the pathology of cystic fibrosis bacterial infection increases *MUC2* and *MUC5AC* mucin mRNA production.

In this report, we present the genomic sequence of 3.55 kb of the *MUC5B* 5'-flanking region that includes the promoter and the first intron. The promoter and first intron were characterised and several important *cis*-elements that bind nuclear factors (Sp1, N-Myc, NF- $\kappa$ B, TGT3, AP-2 $\alpha$  and GATA-1) were identified. *MUC5B* transcriptional activity was studied in colon cancer cell lines that are undifferentiated (HT-29 STD) or that either differentiate in mucus-secreting cells (LS174T) or in enterocytes (Caco-2). It appears that *MUC5B* expression shows cell-specificity and is the highest in mucus-secreting colon cancer cell line LS174T. The roles of the structural and functional properties of *MUC5B* promoter in *MUC5B* gene transcription are discussed.

## **Materials and Methods**

### ***Cloning***

Inserts were prepared using the restriction map of two cosmid clones called ELO9 [21] and BEN1 [8], which cover the 5' proximal region of *MUC5B*. Gel-purified fragments (Qiaquick gel extraction kit, Qiagen) were subcloned into the promoterless pGL3 Basic vector (Promega). Internal deletion mutants were generated by PCR using pairs of primers bearing specific restriction sites at their 5' and 3' ends. PCR products were digested, gel-purified and subcloned into the pGL3 vector that has previously been cut with the same restriction enzymes. All clones were sequenced on both strands on an automatic LI-COR sequencer (ScienceTec, France) using infra-red labelled RV3 and GL2 primers (Promega). Plasmids used for transfection studies were prepared using the Endofree plasmid Mega kit (Qiagen).

### ***Cell culture***

The LS174T cell line was cultured in a 37°C incubator with 5% CO<sub>2</sub> in Dulbecco's modified Eagle's minimum essential medium (DMEM) supplemented with non-essential aa and 10% foetal calf serum (Boehringer-Mannheim). Other human colon adenocarcinoma cell lines were cultured in a 37°C incubator with 10% CO<sub>2</sub> in DMEM supplemented with either 10% (HT-29 STD, HT-29 MTX) or 20% (Caco-2) foetal calf serum. All cell lines were a kind gift from Dr. A. Zweibaum (Unité INSERM N°178, Villejuif, France).

### ***RT-PCR***

Total RNAs from colon cancer cells were prepared using the RNeasy midi-kit from Qiagen. Cells were harvested at 70% of confluence and 1.5 µg of total RNA was used to prepare cDNA (Advantage<sup>TM</sup> RT-for-PCR kit, Clontech). PCR was then performed on 5 µl of

cDNA using specific pairs of primers for *MUC2* and *MUC5B* mucin genes (M.-P. Buisine, N. Moniaux, S. Nollet, E. Destailleur, I. Van Seuningen, A. Laine, N. Porchet and J.-P. Aubert, unpublished work) (*MUC2* forward primer : 5' CTGCACCAAGACCGTCCTCATG 3'; *MUC2* reverse primer : 5' GCAAGGACTGAACAAAGACTCAGAC 3' (position Nt 15291-15312 and Nt 15667-15691, accession number L21998 [22]) ; *MUC5B* forward primer : 5' CTGCGAGACCGAGGTCAACATC 3' ; *MUC5B* reverse primer : 5' TGGGCAGCAGGAGCACGGAG 3' (position Nt 9057-9078 and Nt 10108-10127, accession number Y09788 [9])). Single-stranded oligonucleotides were synthesised by MWG-Biotech, Germany. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as an internal control. PCR reactions were carried out in 50  $\mu$ l final solutions (5  $\mu$ l Perkin-Elmer buffer, 5  $\mu$ l 25 mM MgCl<sub>2</sub>, 5  $\mu$ l 2.5 mM dNTPs, 10 pmoles of each primer, 2U Taq Polymerase (Ampli Taq Gold, Perkin Elmer). Cycling conditions were as follows : 1) Denaturation : 96°C, 10 min for one cycle, 2) Denaturation : 96°C, 30 sec; annealing : 60°C, 30 sec; and extension : 72°C, 1 min for 30 cycles and 3) final extension : 72°C, 15 min. PCR products were analysed on 2% E-gels run in 1X Tris-borate-EDTA buffer (Invitrogen). Mixed ladder was purchased from Invitrogen. PCR products expected sizes are 401 bp for *MUC2* and 415 bp for *MUC5B*.

### ***Primer Extension***

Primer extension reactions were performed using 25  $\mu$ g of total RNA isolated from LS174T and HT-29 MTX cell lines. Primers were end-labelled with [ $\gamma$ -<sup>32</sup>P] dATP. Exon 1 reverse primer : 5'-TGCCTGCGGCACCACGAGCATG-3' (see Fig. 2B) was annealed to RNA at 62°C for 20 min followed by a 10 min incubation at room temperature. Extension was performed as described in the Promega primer extension kit using 200 units of MMLV reverse transcriptase (Advantage<sup>TM</sup> RT-for-PCR kit, Clontech) for 45 min at 42°C. Reactions were stopped by adding 20  $\mu$ l of loading buffer (98% formamide, 10 mM EDTA, 0.1% xylene



cyanol, 0.1% bromophenol blue).  $\phi$ X174 DNA/*Hinf*I dephosphorylated markers were radio-labelled with [ $\gamma$ -<sup>32</sup>P] dATP just before use (Promega). Manual sequencing of fragment 1429 (DNA fragment of *MUC5B* promoter region) was performed using the T7 Sequenase version 2.0 kit (Amersham Inc.). Samples were denatured for 10 min at 90°C before loading on a 6% sequencing gel (Sequagel-6, National diagnostic). The gel was then vacuum-dried and autoradiographed for 3-4 days at -80°C.

### ***Oligonucleotides and DNA fragments***

Oligonucleotides used for PCR are shown in Table I. Those used for gel-shift assays are indicated in Table II. They were synthesised by MWG-Biotech, Germany. Annealed oligonucleotides were radio-labelled using T4 polynucleotide kinase (Promega) and [ $\gamma$ -<sup>32</sup>P] dATP. Radio-labelled probes were separated from free nucleotides on Sephadex G-25 columns (Boehringer-Mannheim). DNA fragment 1472 was produced by PCR (forward primer : 5' CGCGAGCTCGATAACCAAGGGCAGTCA 3', reverse primer : 5' CGCACGCGTTCCACCCCGACCCACAGC 3') and cloned into pGL3 vector. This fragment contains the 8 GA boxes found in intron 1. The fragment was cut out of the vector by a *Sac*I-*Mlu*I double digestion, gel-purified and labelled with [ $\alpha$ -<sup>32</sup>P] dCTP at the 3' end by a fill-in reaction using the Klenow fragment (Boehringer-Mannheim).

### ***Transfections***

Human colon adenocarcinoma cells : HT-29 STD (parental undifferentiated cell line), Caco-2 (enterocytes) and LS174T (goblet cells, mucus-secreting) were electroporated using Gene pulser II apparatus (Biorad) at day 7 (70% confluence). Before electroporation, 25  $\mu$ g of plasmid were added to the cells. In co-transfection experiments, 20  $\mu$ g of pGL3-deletion mutant and 5  $\mu$ g of pCMV-Sp1 or pCMV-Sp3 were used. After a 10 min incubation time at

room temperature, transfected cells were plated and incubated for 48 h at 37°C. Total cell extracts were prepared using 1X Reagent Lysis Buffer (Promega) as described in the manufacturer's instruction manual. Results were corrected for transfection efficiency by co-transfecting 2  $\mu\text{g}$  of pSV- $\beta\text{Gal}$  vector (Promega).  $\beta$ -galactosidase activity was measured in 96 well plates as described in the manufacturer's instruction manual using 50  $\mu\text{l}$  of cell extracts (Promega). Luciferase activity was measured on a Berthold 5901 luminometer on 100  $\mu\text{l}$  of cell extracts using luciferase assay reagent (Promega). The luciferase activity is expressed as fold of induction of the test plasmid activity compared to that of the corresponding control vector (pGL3 control, Promega) after correction for transfection efficiency by dividing by  $\beta$ -galactosidase activity. Each plasmid was assayed in duplicate in at least three separate experiments. pCMV4, pCMV4-Sp1 and pCMV4-Sp3 vectors were a kind gift from Dr. J. M. Horowitz (Raleigh, USA).

### ***Nuclear extract preparation***

Nuclear extracts from cell lines of interest were prepared as described in [23], and kept at -80°C until use. Protein content (5  $\mu\text{l}$  of cell extracts) was measured using the bicinchoninic acid method in 96 well plates as described in the manufacturer's instruction manual (Pierce).

### ***Gel-shift assays***

Nuclear proteins (5  $\mu\text{g}$ ) were pre-incubated for 20 min on ice in 20  $\mu\text{l}$  binding buffer with 2  $\mu\text{g}$  of poly dI-dC (Sigma) and 1  $\mu\text{g}$  sonicated salmon sperm DNA. Radio-labelled DNA probe was added (120,000 c.p.m./reaction) and the reaction left for another 20 min on ice. For super-shift analyses, 1  $\mu\text{l}$  of the antibody of interest (anti-Sp1, anti-AP2 $\alpha$ , anti-GATA-1, anti-c-Myc, anti-N-Myc, anti-Max, anti-c-Fos, anti-NF- $\kappa\text{B}$  p50 and anti-NF- $\kappa\text{B}$  p65, Santa-Cruz laboratories) was added to the proteins and left for 1 h on ice before adding the radio-labelled

probe. Negative controls were carried out using 1  $\mu$ l of irrelevant antibody in the reaction mixture. Competitions were performed with 50, 150, and 300 excess of cold oligonucleotide in the reaction mixture before adding the radio-labelled probe. Commercial Sp1 oligonucleotide and HeLa cell nuclear extract used as positive controls were from Promega. Reaction was stopped by adding 2  $\mu$ l of loading buffer, loaded onto a 4% non-denaturing polyacrylamide gel (for oligonucleotides) and electrophoresis conditions were as described in [24]. Gels were pre-run for 30 min at 200V before loading the samples and running the electrophoresis at 200V for 2-3 h in 0.5X Tris-borate-EDTA buffer at 4°C. For DNA probes longer than 250 bp (1472), 6% gels were run as described in [25] with slight modifications : pre-run was 1 h 15 min long and electrophoresis was run for 4 h 15 min at 4°C. Gels were then vacuum-dried and autoradiographed overnight at -80°C.

### ***Southwestern and immunoblotting***

25  $\mu$ g of nuclear proteins were resolved on a 10% SDS-gel and then electro-transferred onto a nitrocellulose membrane (0.22  $\mu$ m, Schleicher and Schuell). Proteins were submitted to a denaturation/renaturation cycle as described by Jackson [26] before radio-labelled probe ( $1 \times 10^6$  c.p.m./ml of buffer) was added to the mixture and left for 30 min at room temperature. When necessary, cold competitor (150 fold excess) was pre-incubated with the membrane before adding the radio-labelled probe. The membrane was then washed and autoradiographed overnight at -80°C. The radio-labelled probe was washed off by incubating the membrane twice for 5 min in boiling 0.1X SSC, 0.1% SDS solution and rinsed once with 3X SSC. Immunostaining was then performed using the anti-Sp1 antibody PEP2 (Santa-Cruz laboratories) at a 1:1,000 dilution. Alkaline phosphatase-conjugated goat anti-rabbit antibody (Promega) was used at a 1:4,725 dilution. Protein bands were developed using nitroblue tetrazolium chloride and 5-bromo-4-chloro-3-indolylphosphate *p*-toluidine salt (Life

Technologies Inc.) as chromogenic substrates for alkaline phosphatase. Prestained protein molecular weight standards were from Life Technologies Inc.

*DNA sequence and transcription factor binding site analyses*

DNA sequences were analysed using PC-GENE software and the TRANSFAC database was used to define potential transcription factor binding sites within the clones of interest. The search was conducted using MatInspector V2.2 software [27].

## Results

### *DNA sequence and characterisation of MUC5B 5'-flanking region*

Two genomic clones, ELO9 and BEN1, that we previously isolated to study *MUC5B* and *MUC5AC* gene organisation [9,21], were used in this work to characterise the 5'-flanking region of *MUC5B*. Here, we report the genomic sequence obtained and an analysis of transcriptional activity of the 5'-flanking region including 956 bases of DNA sequence located upstream of the transcriptional start site as well as 2637 bases downstream that include the first two exons and the first intron. Between the transcription start site and the ATG codon coding the first methionine, lays the 5' UTR long of 56 nucleotides (Fig. 1A). The first exon encodes the first 23 aa (MGAPSACRTLVLALAAMLVVPQA) and is separated from exon 2, which is composed of 19 aa (ETQGPVEPSWGNAGHTMDG), by an intron long of 2511 nucleotides (Fig. 1B).

Analysis of the DNA sequence using PC-GENE software and TRANSFAC database [27] showed that the 5'-flanking region upstream of the transcription start site contains a TATA box-like sequence TACATAA located at bases -32/-26 upstream of the transcription start site (Fig. 1A). In the close vicinity to the TATA box-like sequence were found potential binding sites for c-Myc/N-Myc/Max transcription factors (consensus CACGTG), Sp1, and numerous GC and CACCC boxes. It is of interest to note that GC and CACCC boxes are always clustered. Further upstream, putative binding sites were found for AP-1, NF- $\kappa$ B, CCAAT, CREB and glucocorticoid receptor element (GRE) transcription factors as well as for a silencer called CIIS1. A striking feature was the presence, throughout the 5'-flanking region upstream of the TATA box, of numerous putative binding sites for transcription factors known to be involved in intestinal (HNF-1, HNF-3, TGT3, GKLF) or respiratory (HNF-3,

TTF-1) cell differentiation. Numerous *Hpa*II sites that could be potential methylation sites were found, two of them being located in the 5'UTR region.

In the first intron (Fig. 1B), a high amount of GC boxes, Sp1 binding sites and CACCC boxes are present. More importantly were found clustered in the central part of the intron a CACCC box followed by an array of 8 GA boxes repeated in tandem containing the consensus sequence GGGGAGGGGCT each separated by 8-10 bp. Other putative sites for known transcription factors are shown in Fig. 1B and include three canonical Sp1 binding sites, two AP-2 sites, one NF- $\kappa$ B site, one GATA-1 site and one Adh-1 site.

### ***Characterisation of the MUC5B transcription start site***

Primer extension reactions were performed using RNAs from the two mucus-secreting colon cancer cell lines LS174T and HT-29 MTX to identify the transcription start site in *MUC5B* 5'-flanking region. These two cell lines were chosen as they both express *MUC5B*. The result, shown in Fig. 2A, indicates that the reactions carried out with reverse primer E1 localised in exon 1 (sequence underlined in Fig. 2B) produced one extension product of the same size (126 bp) in both cell lines. In the sequence, the start site matches a cytosine residue that we numbered +1. The 5'UTR region is 56 nucleotides long (Fig. 2B).

### ***MUC5B gene expression in LS174T, Caco-2 and HT-29 STD colon cancer cells***

To investigate the cell-specificity expression of *MUC5B*, three colon cancer cell lines showing different phenotypes were chosen : the mucus-secreting LS174T, the Caco-2 enterocyte and the undifferentiated HT-29 STD cell lines. *MUC2* expression was studied in parallel as a positive control of RT-PCR as it is the main mucin expressed in colon. *MUC5B* (415 bp PCR product) and *MUC2* (401 bp PCR product) mRNA expression was studied by reverse transcription-PCR (RT-PCR) from 70% confluent cells in order to correlate RT-PCR

results with transfection experiments. As shown in Fig. 3, undifferentiated cell line HT-29 STD slightly expresses *MUC2* and *MUC5B* mRNAs. Caco-2 cells that spontaneously differentiate in enterocytes also show a very low expression of the two genes. Overall, in these two cell lines, mucin gene expression remains very discrete when compared with GAPDH internal control. On the other hand, *MUC2* and *MUC5B* are highly expressed in the mucus-secreting cell line LS174T. The level of expression of both genes is maximal after cells have differentiated into mucus-secreting cells. Moreover, one can notice that *MUC2* and *MUC5B* follow the same pattern of expression in the three colon cancer cell lines studied.

### ***Promoter activity of the MUC5B 5'-flanking region***

In order to identify DNA sequences involved in *MUC5B* transcriptional regulation, constructs were generated in the promoterless pGL3 Basic vector and analysed for transcriptional activity after cell electroporation. Insert sequences were confirmed by infra-red sequencing of both strands and aligned with ELO9 and BEN1 sequences.

Transfection experiments of deletion fragments of *MUC5B* promoter were carried out in Caco-2, LS174T and HT-29 STD cell lines. Seven fragments were generated that cover 956 nucleotides upstream of the transcription start site (Fig. 4A). The fragments 1916, 1896, 1597 and 1596 contain the TATA box-like sequence (TACATAA), the three Sp1 binding sites and the CACCC box. In fragments 1916 and 1597, the 5'UTR region was included. In LS174T cells, the *MUC5B* 5'-flanking region shows a highly active promoter activity. In these cells, the fragment 1896 that covers the whole region shows a lower luciferase activity when compared to the region restricted to the first 223 nucleotides upstream of the transcription start site (1596) (Fig. 4B). Interestingly enough, when the 5'UTR region was included in the fragments corresponding to the whole region (1916) or to the 223 first bases (1597), luciferase activity dropped by approximately 2 fold. This result indicates that the 5'UTR region of

*MUC5B* plays a role in regulating promoter activity. In conclusion, the first 223 bp (fragment 1596) adjoining the transcription start site suffice to drive basal promoter activity of the luciferase reporter gene. Three other deletion mutants were then constructed in order to identify a potential inhibitory domain in the promoter that would explain the low luciferase activity of the fragment 1896. Fragment 1895 covers the upstream 734 bases (-956/-223) and fragments 1595 and 1598 include bases -446/-223 and -956/-441, respectively. Experiments show that in all cell lines 1595 has no activity whereas fragments 1895 and 1598 act as inhibitory domains. Thus, this result confirms the fact that inhibitory factors bind to *cis*-elements present in the -956/-441 region of the promoter. Finally, *MUC5B* promoter activity is minimal in Caco-2 and HT-29 STD cell lines when compared to that in LS174T cells.

#### ***Binding studies of MUC5B promoter with nuclear proteins***

In order to characterise *cis*-elements and *trans*-nuclear factors that could account for *MUC5B* transcription regulation in colon cancer cells, electromobility shift assays were carried out with double-stranded DNA probes containing binding sites for specific nuclear factors (Fig. 5). In panel A, two retarded bands (1 and 2) were obtained when Caco-2 nuclear extracts were incubated with the double-stranded oligonucleotide comprising the c-Myc/N-Myc/Max putative binding site (lane 2). Addition of c-Myc or N-Myc antibodies prior to the radio-labelled probe dramatically reduced the intensity of the complex #1 but had no effect on complex #2 (lanes 3-4). On the other hand, addition of the Max antibody had no effect (lane 5). Complex #1 was not visualised in LS174T cells (lanes 7-10). Sp1 binds to the CACCC box in both cell lines (complex #3) as addition of Sp1 antibody resulted in a totally super-shifted complex (band #4) (lanes 11-14).

In panel B, is shown the binding of NF- $\kappa$ B (lanes 1-7). When Caco-2 nuclear proteins were incubated with the double-stranded oligonucleotide containing the putative NF- $\kappa$ B



binding site (lane 2), three retarded bands were visualised on the autoradiogram (bands 1-3). When incubated with LS174T nuclear proteins (lane 5) only bands #1 and #3 could be seen. Upon addition of the specific NF- $\kappa$ B p50 antibody (lane 3), complexes #2 and #3 were abolished. With the antibody directed against the NF- $\kappa$ B p65 sub-unit, only complex #2 disappeared. The complex #1 is therefore due to an interaction with another nuclear protein since the addition of these two antibodies did not have any effect on the shifted band. Binding studies with the TGT3 putative binding site (lanes 8-10) resulted into two high mobility retarded bands with Caco-2 nuclear proteins (lane 9, complexes #5 and #6) but only one low mobility shifted band with LS174T nuclear extracts (lane 10, complex #4).

In panel C, studies with the CCAAT DNA probe showed one retarded band with Caco-2 nuclear extracts (lane 2, complex #1), two more retarded bands with a higher mobility (complexes #2 and #3) were visualised when incubated with LS174T nuclear extracts (lane 3).

### ***Binding studies of MUC5B first intron with nuclear proteins***

Given that first introns have been shown to play roles in the regulation of tissue-specific intestinal genes, we looked whether such potential DNA binding *cis*-elements could be found in *MUC5B* rather large first intron (2.5 kb). A highly striking feature was the presence of 8 clustered GA boxes repeated in tandem located nearby a CACCC box. All these *cis*-elements represent potential binding sites for the ubiquitous nuclear factor Sp1. Gel-shift assay experiments were performed with nuclear extracts from the three cell lines of interest. As they gave the same results only those with LS174T nuclear proteins are shown. When we used double stranded DNA representative of the CACCC box one low mobility retarded band was visualised (complex #1) (Fig. 6A, lane 2) that was totally super-shifted when an anti-Sp1 antibody was added in the reaction mixture (lane 6). The same result was obtained with the double-stranded oligonucleotide containing the GA box consensus sequence

GGGGAGGGGCT (lanes 3, 7). A similar experiment conducted with an oligonucleotide containing a Sp1 found in the 3' region of the intron did not produce any retarded bands (lanes 4 and 8).

In order to confirm the binding of Sp1 to the GA boxes, southwestern- and immunoblotting studies were performed (Fig. 6B). The nuclear protein recognised by the radioactive GA box DNA probe shows a  $M_r$  around  $95 \times 10^3$  kDa in the three cell lines (lanes 7-9). When cold competition was performed with an excess of the cold GA box DNA probe ( $\times 150$ ) the signal was almost totally abolished, demonstrating the specificity of the interaction (lanes 10-12). As a positive control for Sp1 binding, the blot was then probed with a DNA probe representative of one of the Sp1 binding sites present in *MUC5B* promoter (see fig. 5, lanes 11-14). The same band was visualised (lanes 1-3) and cold competition also resulted in the loss of the signal (lanes 4-6). The same blot was then probed with the anti-Sp1 antibody (lanes 13-15). The band visualised after immunostaining superimposed with the radioactive band obtained from the southwestern-blotting study, confirming the fact that Sp1 binds to the GA box. We can conclude from these results that the Sp1 nuclear factor engages with the CACCC and GA boxes found in *MUC5B* intron 1. It is likely that the other 7 GA boxes that contain the same consensus sequence GGGGAGGGGCT also interact with Sp1.

To verify this hypothesis, gel-shift experiments were performed using the fragment *1472* that contains the eight GA boxes as the DNA probe. Fig. 6C shows that the shifted band obtained when *1472* was incubated with nuclear proteins from LS174T cells (lane 2, complex #1) was completely super-shifted when the anti-Sp1 antibody was added to the reaction mixture (lane 3, complex #2). The shift was shown to be specific by performing competition experiments with cold DNA fragment *1472* (lanes 4-6) or cold GA box oligonucleotide (lanes 7-9) as competitors. The shift is almost totally inhibited when competitors are added at a molar ratio of 150 (lanes 5 and 8) and totally disappears at a ratio of 300 (lanes 6 and 9).

Using mutated oligonucleotides with a modified GA box sequence (see table II) only produced a partial inhibition of the retarded band (lanes 10-12).

Among the other predicted binding sites present in intron 1, we studied NF- $\kappa$ B, AP-2 and GATA-1 binding (Fig. 7). Two shifted bands were obtained when Caco-2 nuclear proteins were incubated with the AP-2 oligonucleotide (lane 2). Out of these two bands, the upper one disappeared upon addition of the anti-AP-2 $\alpha$  antibody (lane 3). The specificity of the binding was confirmed by the absence of any effect when an irrelevant anti-c-Fos antibody was used (lane 4). The complex #1 was not seen in LS174T cells (not shown). When we used the double-stranded GATA oligonucleotide, one shifted band (complex #3) was visualised with Caco-2 nuclear proteins (lane 6). The intensity of the shifted band decreased substantially when an anti-GATA-1 antibody was added to the reaction mixture (lane 5) whereas there was no effect with the anti-Sp1 antibody (lane 8). No shift was visualised with the NF- $\kappa$ B oligonucleotide in either cell line (not shown).

### ***Sp1 trans-activation of MUC5B promoter***

In order to demonstrate the active participation of Sp1 in *MUC5B* promoter activity, co-transfection experiments were performed in the three cell lines of interest. Sp3, another member of the Sp family known to compete with Sp1 for the same binding sites was also studied (Fig. 8). Co-transfection of Sp1 with fragment 1896 (-956/-1) led to an increase in luciferase activity in LS174T and Caco-2 cells (2 and 3.5 fold, respectively) whereas co-transfection with Sp3 led to a slight decrease of activity in LS174T cells and had no effect in Caco-2 cells. When the same experiments were conducted with the fragment 1596 (-223/-1) no significant effect was observed in either cell line. In HT-29 STD cells, co-transfection with Sp1 had no effect whereas co-transfection with Sp3 dramatically decreased the luciferase activity of both fragments by about five fold.

## Discussion

Molecular mechanisms involved in the control of gene expression in intestinal cells are still poorly understood. An understanding of the regulatory network of nuclear proteins that direct transcriptional initiation of genes will provide insight into the mechanisms of normal intestinal development and differentiation as well as disease processes such as neoplasia [28]. Since mucins are the main components of human epithelial secretions, responsible for the mucus rheology and physical properties, and since mucin genes display altered expression in neoplastic and preneoplastic processes affecting the colon [3,29], it is essential to understand the molecular mechanisms that prevail to their expression. Thus, characterisation of the promoter regions as well as the nuclear factors involved in the regulation of mucin gene expression is necessary.

In order to identify regulatory regions involved in *MUC5B* transcription, *MUC5B* promoter region as well as the first intron have been sequenced and analysed. Transcriptional activity was studied in three colon cancer cell lines and transcription factor binding to *cis*-elements present either in the promoter or the first intron were identified. The 5'-flanking region upstream of the transcriptional start site revealed the presence of a TATA-like sequence TACATAA at -26/-32, which is a characteristic of tissue-specific genes. *MUC5B* promoter contains a high amount of GC and CACCC boxes that are known to bind Sp1. In this report, Sp1 binding was demonstrated to occur in this region in the three cell lines. Involvement of Sp1 in regulation of mucin genes has already been shown for *MUC1* [30] and *MUC2* [13,14]. In this report, the transactivating effect of Sp1 on *MUC5B* promoter was shown to be effective on the whole promoter region but not on the fragment containing the first 223 nucleotides adjacent to the TATA box. This result may be explained by the fact that other transcription factors or epigenetic mechanisms like methylation are required to mediate

Sp1 activity. In *MUC5B* promoter, Sp1 binding sites were often found to be nearby binding sites for other specific factors. This phenomenon usually indicates that the clustered factors act then in conjunction with each other to modulate transcription [31]. Among the other factors, NF- $\kappa$ B may be a good candidate as it has already been shown to increase *MUC2* promoter activity in epithelial cells via a Src-dependent pathway in response to bacterial exoproducts [18]. *MUC5B* promoter contains a  $\kappa$ B element that binds NF- $\kappa$ B in Caco-2 cells. In these cells, N-Myc, c-Myc as well as AP-2 and GATA factors may also control regulation of *MUC5B* since these factors were shown in this report to bind to *MUC5B* promoter and first intron, respectively.

The first intron of *MUC5B*, whose sequence is described for the first time in this report is very large (2.5 kb) and contains a striking feature, that is a central domain long of 259 bp, which contains 8 GA boxes repeated in tandem. Sp1 was shown to bind throughout this domain suggesting that this domain may serve as a docking platform for Sp1. Such kind of repetitive sequences located in introns of the 5'-flanking regions have already been described for human immunoglobulin heavy chain, insulin, or serotonin transporter genes. This type of sequence remains however relatively rare in the human genome [32]. They are usually rich in GC and GA residues and induce looping of DNA therefore their presence is often associated with a role in regulating gene transcription [32]. Such a role may thus be played by the first intron of *MUC5B*.

*MUC5B* promoter activity was studied in three intestinal cancer cell lines showing different phenotypes in order to address cell-specific activity. The cells were either undifferentiated (HT-29 STD), differentiated in enterocytes (Caco-2) or mucus-secreting (LS174T). Clearly, *MUC5B* promoter is not active in undifferentiated and enterocytes colon cancer cells. It is however very active in LS174T mucus-secreting cells. This result confirmed our RT-PCR results in which *MUC5B* mRNA expression was important in LS174T cells and

almost absent from HT-29 STD and Caco-2 cells. Our results regarding *MUC2* and *MUC5B* expression in LS174T cells are in good agreement with those published by Van Klinken *et al.* [33]. A recent study by these authors also showed that *MUC2* and *MUC5B* were expressed in the same subset of colonic cells that is goblet cells in the crypt base; co-localised expression was also shown in the small granules within goblet cells [6]. Why does the cell need to simultaneously express several mucin genes is another issue that will have to be addressed. Furthermore, the knowledge of the mechanisms that control transcription of the four human mucin genes *MUC6*, *MUC2*, *MUC5AC* and *MUC5B* clustered on the chromosome 11 will certainly help in understanding the role of some regulatory factors in intestinal diseases in which altered expression of mucin genes has already been demonstrated [34].

Numerous putative binding sites for transcription factors involved in intestinal (HNF-1, HNF-3, three consensus binding sites for TGT3, GKLf) or respiratory (10 consensus binding sites for TTF-1) cell differentiation were found throughout *MUC5B* promoter. *In situ* hybridisation studies performed in our laboratory showing early expression of human mucin genes during embryonic and foetal development in intestine and respiratory tracts suggest the involvement of these genes during ontogenesis [35,36]. Identification of silencers is also important to elucidate *MUC5B* down-regulation in certain cells. A putative binding site for such a factor, called CHS1 [37], was found in *MUC5B* promoter in fragments that showed inhibitory activity (1895 and 1598).

In conclusion, we have characterised the 5'-flanking region of *MUC5B* and showed that promoter activity was the highest in mucus-secreting cells. Important features and *cis*-elements that bind specific transcription factors were identified in the promoter (TATA box, Sp1, NF- $\kappa$ B, c-Myc, N-Myc) and the first intron (Sp1, GATA-1 and AP-2 $\alpha$ ) of *MUC5B* that might account for the cell-specific expression pattern of that gene. We also showed that Sp1 is involved in the regulation of *MUC5B* promoter activity. Experiments are now in progress to

identify transcription factors and cell signalling pathways responsible for the altered expression of *MUC5B* in intestinal diseases. Of particular interest, will be the deciphering of the molecular mechanisms implying *MUC5B* gene in gastro-intestinal cell differentiation. Moreover, experiments are now in progress in the laboratory to determine the influence of intron 1 and intron 37, that has already been shown to bind Sp1 and NF1-*MUC5B* nuclear factors [12], on *MUC5B* promoter activity.

Acknowledgements : We are indebted to Danièle Petitprez, Pascal Mathon, Marie-Paule Delescaut and Isabelle Clavereau for their excellent technical help. We thank Dr. J. M. Horowitz for the kind gift of pCMV4, pCMV4-Sp1 and pCMV4-Sp3 vectors. We thank the members of our E.U. consortium CEEBMH4-3222 for stimulating discussions. This work was supported by grants from l'Association de Recherche contre le Cancer and le Comité du Nord de la Ligue Nationale contre le Cancer.

## References

1. Gendler, S. J. and Spicer, A. P. (1995) *Ann. Rev. Physiol.* **57**, 607-634
2. Pigny, P., Guyonnet-Dupérat, V., Hill, A.S., Pratt, W.S., Galiegue-Zoutina, S., Collyn D'Hooge, M., Laine, A., Van Seuning, I., Degand, P., Gum, J.R., Kim, Y.S., Swallow, D.M., Aubert, J.-P. and Porchet, N. (1996) *Genomics* **38**, 340-352
3. Lesuffleur, T., Zweibaum, A. and Real, F. X. (1994) *Crit. Rev. Hematol. Oncol.* **17**, 153-180
4. Aubert, J.-P., Porchet, N., Crépin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D. and Degand P. (1991) *Am. J. Respir. Cell Mol. Biol.* **5**, 178-185
5. Vandehaute, B., Buisine, M.-P., Debailleul, V., Clément, B., Moniaux, N., Dieu, M.-C., Degand, P., Porchet, N. and Aubert, J.-P. (1997) *J. Hepatol.* **27**, 1057-1066
6. Van Klinken, B. J.-W., Dekker, J., Van Gool, S. A., Van Marle, J., Büller, H. A. and Einerhand, A. W. C. (1998) *Am. J. Physiol.* **274**, G871-G878
7. 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



8. Desseyn, J.-L., Guyonnet-Dupérat, V., Porchet, N., Aubert, J.-P. and Laine, A. (1997) *J. Biol. Chem.* **272**, 3168-3178
9. Desseyn, J.-L., Aubert, J.-P., Van Seuningen, I., Porchet, N. and Laine, A. (1997) *J. Biol. Chem.* **272**, 16873-16883
10. Desseyn, J.-L., Buisine, M.-P., Porchet, N., Aubert, J.-P. and Laine, A. (1998) *J. Biol. Chem.* **273**, 30157-30164
11. Offner, G., Nunes, D. P., Keates, A. C., Afdhal, N. H. and Troxler, R. F. (1998) *Biochem. Biophys. Res. Comm.* **251**, 350-355
12. Pigny, P., Van Seuningen, I., Desseyn, J.-L., Nollet, S., Porchet, N., Laine, A. and Aubert, J.-P. (1996) *Biochem. Biophys. Res. Comm.* **220**, 186-191
13. Velcich, A., Palumbo, L., Selleri, L., Evans, G. and Augenlicht, L. (1997) *J. Biol. Chem.* **272**, 7968-7976
14. Gum, J. R., Hicks, J. W. and Kim, Y. S. (1997) *Biochem. J.* **325**, 259-267
15. Gum, J. R., Hicks, J. W., Gillespie, A.-M., Carlson, E. J., Komuves, L., Karnik, S., Hong, J. C., Epstein, C. J. and Kim, Y. S. (1999) *Am. J. Physiol.* **276**, G666-G676
16. Nogami, H., Ohmori, H., Li, J.-D., Gallup, M., Gum, J., Kim, Y. and Basbaum, C. (1997) *Gene* **198**, 191-201

17. Li, J.-D., Dohrman, A. F., Gallup, M., Miyata, S., Gum, J. R., Kim, Y. S., Nadel, J. A., Prince, A. and Basbaum, C. (1997) *Proc. Natl. Acad. Sci.* **94**, 967-972
18. Li, J.-D., Feng, W., Gallup, M., Kim, J.-H., Gum, J., Kim, Y. and Basbaum C. (1998) *Proc. Natl. Acad. Sci.* **95**, 5718-5723
19. Li, D., Gallup, M., Fan, N., Szymkowski, D. E. and Basbaum C. (1998) *J. Biol. Chem.* **273**, 6812-6820
20. Dohrman, A., Miyata, S., Gallup, M., Li, J.-D., Chapelin, C., Coste, A., Escudier, E., Nadel, J. and Basbaum, C. (1998) *Biochem. Biophys. Acta* **1406**, 251-259
21. Buisine, M.-P., Desseyn, J.-L., Porchet, N., Degand, P., Laine, A. and Aubert, J.-P. (1998) *Biochem. J.* **332**, 1-10
22. Gum, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) *J. Biol. Chem.* **269**, 2440-2446
23. Van Seuningen, I., Ostrowski, J., Bustelo, X. R., Sleath, P. and Bomsztyk, K. (1995) *J. Biol. Chem.* **270**, 26976-26985
24. Vindevoghel, L., Chung, K.-Y., Davis, A., Kouba, D., Kivirikko, S., Alder, H., Uitto, J. and Mauviel, A. (1997) *J. Biol. Chem.* **272**, 10196-10204

25. Taylor, J. D., Badcoe, I. G., Clarke, A.R. and Halford, S.E. (1991) *Biochemistry* **30**, 8743-8753
26. Jackson, S. P. (1993) In *Gene transcription, A practical approach*, (Hames, D., and Higgins, S. J., eds.), pp. 189-242, IRL Press, Oxford.
27. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) *Nucl. Acid Res.* **23**, 4878-4884
28. Traber, P. G. (1998) *Phil. Trans. R. Soc. Lond.* **353**, 911-914
29. 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) *Gastroenterology* **110**, 84-91
30. Kovarik, A., Lu, P. J., Peat, N., Morris, J. and Taylor-Papadimitriou, J. (1996) *J. Biol. Chem.* **271**, 18140-18147
31. Kadonaga, J. T., Carner, K. R., Marsiarz, F. R. and Tjian, R. (1987) *Cell* **51**, 1079-1090
32. Nakamura, Y., Koyama, K. and Matsushima, M. (1998) *J. Hum. Genet.* **43**, 149-152
33. Van Klinken, B. J. W. E., Oussoren, E., Weeninck, J. J., Strous, G. J., Buller, H. A., Dekker, J. and Einerhand, A. W. C. (1996) *Glycoconj. J.* **13**, 757-768

34. Lesuffleur, T., Porchet, N., Aubert, J.-P., Swallow, D., Gum, J. R., Kim, Y. S., Real, F. X. and Zweibaum, A. (1993) *J. Cell Sci.* **106**, 771-783
35. Buisine, M.-P., Devisme, L., Savidge, T. C., Gespach, C., Gosselin, B., Porchet, N. and Aubert, J.-P. (1998) *Gut* **43**, 519-524
36. Buisine, M.-P., Devisme, L., Copin, M.-C., Durand-Réville, M., Gosselin, B., Aubert, J.-P. and Porchet, N. (1999) *Am. J. Respir. Cell Mol. Biol.* **20**, 209-218
37. Savagner, P., Miyashita, T. and Yamada, Y. (1990) *J. Biol. Chem.* **265**, 6669-6674

pGL3 deletion mutant	Position of the DNA fragment in <i>MUC5B</i> 5'-flanking region	Orientation S : sense, AS :	5' ← 3' sequences of the oligonucleotides used in PCR reactions
1916	-956/+57	S	CGC <u><i>GGTACCG</i></u> GAGGTATTGCAGCGC
		AS	CGC <u><i>ACGCGT</i></u> CCTGGGGGCTGGCAC
1896	-956/-1	S	CGC <u><i>GGTACCG</i></u> GAGGTATTGCAGCGC
		AS	CGC <u><i>ACGCGT</i></u> TGCTTGCTCCCCTGGGGGC
1597	-223/+57	S	CGC <u><i>GAGCTC</i></u> TCCCTTCCGATTCTCAGA
		AS	CGC <u><i>ACGCGT</i></u> CCTGGGGGCTGGCAC
1596	-223/-1	S	CGC <u><i>GAGCTC</i></u> TCCCTTCCGATTCTCAGA
		AS	CGC <u><i>ACGCGT</i></u> TGCTTGCTCCCCTGGGGGC
1895	-956/-223	S	CGC <u><i>GGTACCG</i></u> GAGGTATTGCAGCGC
		AS	CGC <u><i>ACGCGT</i></u> AGAGAAGGGCCTACCCCA
1595	-446/-223	S	CGC <u><i>GAGCTC</i></u> CAGCCTCGTGGGCCCTCC
		AS	CGC <u><i>ACGCGT</i></u> AGAGAAGGGCCTACCCCA
1598	-956/-441	S	CGC <u><i>GGTACCG</i></u> GAGGTATTGCAGCGC
		AS	CGC <u><i>GAGCTC</i></u> CAGGCATGCACATGGC

**Table I** : Sequences of the pairs of oligonucleotides used in PCR to produce deletion mutants in *MUC5B* 5'-flanking region (see Fig. 1). *Kpn*I (GGTACC), *Sac*I (GAGCTC), and *Mlu*I (ACGCGT) sites were added at the end of the primers to direct subcloning, they are italicised and underlined. Positions of the DNA fragments relative to the transcription start site are indicated.

Oligonucleotide name (sense orientation)	Sequence (5' → 3')
c-Myc/N-Myc/Max site/promoter	GAGAGCAAACACACGTGGCTGGAG
CACCC/Sp1 site/promoter	CTGCTGGGGGTGGGGGCGTCCTGG
NF-κB site/promoter	CAGAGGCTGGGATTCCCTCCAGC
CCAAT site/promoter	GGCTGGCATTGGACAAGCGCACGT
TGT3 site/promoter	GAGGACCCCGGGCAGGTCCAGGAC
CACCC box/intron 1	TTAGGGGACACACCCCAAAGT
GA box/intron 1	GCAGGGGAGGGGCTGCCTGGG
Mutated GA box #1	GC <b><i>ATT</i></b> AGAGGGGCTGCCTGGG
Mutated GA box #2	GCAGGG <b><i>ATT</i></b> GGGCTGCCTGGG
Mutated GA box #3	GCAGGGGAG <b><i>CT</i></b> ACTGCCTGGG
Sp1 site/intron 1	CCCTTGGGGCGGCCAACGCCT
AP-2 site/intron 1	ACGTCCTGTGGCCGGGGGAAGGCA
GATA-1 site/intron 1	ACGGAGCAGATAACCAAGGGC
NF-κB site/intron 1	GCAGGGGGCTTCCCCCAGTGG

**Table II** : Sequences of the sense oligonucleotides used for gel-shift assay experiments. Mutated bases are in bold and italicised. Antisense complementary oligonucleotides were also synthesised and annealed to the sense oligonucleotide to produce double-stranded DNA.

## Figure legends

**Figure 1 : DNA sequence of the human *MUC5B* promoter (A) and intron 1 (B).** (A) The TATA-like box (TACATAA) at -32/-26 is double-underlined and potential binding sites for known transcription factors are in grey. The transcription start site is designated as +1 and first ATG is bold and italicised. *HpaII* (C\*Cgg) potential methylation sites are also indicated. (B) DNA and peptide sequences of exon 1 and exon 2 surrounding intron 1 are bold and italicised. Brackets indicate the beginning and end of intron 1 sequence. DNA sequences used for binding assays are boxed.

**Figure 2 : Transcription start site of human *MUC5B* mucin gene.** Schematic representation of the 5'-flanking region of *MUC5B*. The bent arrow indicates the transcription start site position (upper panel). (A) : Primer extension reactions were performed on 25  $\mu$ g of total RNA from LS174T and HT-29 MTX cell lines. A product of 126 bp was produced when using the exon1 reverse primer.  $\phi$ X174 DNA/*HinfI* dephosphorylated markers previously radio-labelled and denatured are on the left hand side of the gel. (B) : DNA sequence surrounding the transcription start site. The reverse exon1 oligonucleotide used for primer extension experiments was designed from the underlined sequence.

**Figure 3 : Mucin gene expression profiles in colon cancer cells by RT-PCR.** RT-PCR was performed on total RNA extracted from cells at day 7. Specific oligonucleotides for *MUC2* and *MUC5B* generate 401 and 415 bp PCR products, respectively. *GAPDH* PCR product is 980 bp long. PCR products were separated on a 2% agarose gel. Molecular mass markers are on the right hand side.

**Figure 4 : Position and transcriptional activity of the pGL3 deletion mutants in *MUC5B* promoter.** (A) : Schematic representation of the localisation of the different pGL3 deletion mutants covering 1013 nucleotides upstream of exon 1. Numbering refers to the transcription start site designated +1. The TATA box location in fragments 1916, 1896, 1597, and 1596 is indicated. (B) : transcriptional activity of the deletion mutants was studied in enterocytic Caco-2 (white), in mucus-secreting LS174T (black) and in undifferentiated HT-29 STD (grey) colon cancer cell lines. Background activity of pGL3 Basic promoterless vector used to subclone *MUC5B* fragments is shown. The results are means +/- S.D. and represent more than three different experiments in duplicate for each fragment.

**Figure 5 : Binding of nuclear factors to the promoter of *MUC5B*.** Autoradiograms of the gel-shift assays in which 5  $\mu$ g of LS174T (*LS174T*) or Caco-2 (*Caco-2*) nuclear extracts were incubated with double-stranded radio-labelled oligonucleotides representative of the c-Myc/N-Myc/Max (-108/-98, panel A, lanes 2-5 and 7-10), CACCC/Sp1 (-196/-185, panel A, lanes 11-14), NF- $\kappa$ B (-372/-364, panel B, lanes 2-7), TGT3 (-791/-785, panel B, lanes 9 and 10) and CCAAT (-662/-658, panel C, lanes 2 and 3) binding sites found in *MUC5B* promoter. (A) : No proteins added (lanes 1, 6, 15). Nuclear extracts were either incubated in the absence (-) (lanes 2, 7, 11, 13) or the presence (+) (lanes 3-5, 8-10, 12, 14) of 1  $\mu$ l of the antibody of interest. (B) : No proteins added (lanes 1, 8). Nuclear extracts were either incubated in the absence (-) (lanes 2, 5, 9, 10) or the presence (+) (lanes 3, 4, 6, 7) of 1  $\mu$ l of the antibody of interest. (C) : No proteins added (lane 1). The CCAAT radio-labelled probe was incubated with Caco-2 (lane 2) or LS174T (lane 3) nuclear extracts. DNA-protein complexes are indicated by a number on each side of the autoradiograms.



**Figure 6 : Sp1 binds to the 8 clustered GA boxes and the CACCC box found in intron 1.**

(A) : autoradiogram of the gel-shift assay in which 5  $\mu$ g of LS174T nuclear extracts were incubated with double-stranded oligonucleotides representative of the CACCC (lane 2), GA (lane 3), Sp1 box (lane 4) found in *MUC5B* intron 1 and with the GC box located in intron 37 [12] (lane 5). Supershift analyses were performed by adding an antibody directed against Sp1 (lanes 6-10 and 12). A positive control was included by performing shift analysis with a commercial Sp1 oligonucleotide mixed with LS174T extracts (lane 10) and Hela cell extracts (lane 11). Sp1 supershift is shown with Hela cells in lane 12. The arrow 1 indicates the position of the DNA-protein complex and the arrow 2 the position of the supershifted complex when the antibody of interest was added. (B) : Autoradiogram of the southwestern blot of 25  $\mu$ g of nuclear proteins from Caco-2 (lanes 2, 5, 8, 11, 14), LS174T (lanes 3, 6, 9, 12, 15) and HT-29 STD (lanes 1, 4, 7, 10, 13) cells incubated with the GC (lanes 1-6) and GA (lanes 7-12) boxes radio-labelled probes. Cold competition with 150 fold excess of GC (lanes 4-6) and GA (lanes 10-12) boxes. Sp1 immunostaining of the same blot with a commercial antibody directed against Sp1 (PEP2) (lane 13-15). (C) : autoradiogram of the gel-shift assay with *I472* as the probe with 5  $\mu$ g of LS174T nuclear proteins. No extracts (lane 1), proteins and *I472* probe (lane 2), super-shift with anti-Sp1 antibody (lane 3). Competition with cold *I472* DNA, molar ratio x50 (lane 4), x150 (lane 5) and x300 (lane 6). Competition with cold GA box oligonucleotide, molar ratio x50 (lane 7), x150 (lane 8) and x300 (lane 9). Competition with cold mutated GA box oligonucleotides with a molar ratio x 150, \*GA box #1 (lane 10), \*GA box #2 (lane 11), and \*GA box #3 (lane 12). The arrow 1 indicates the position of the DNA-protein complex and the arrow 2 the position of the supershifted complex when the anti-Sp1 antibody was added.

**Figure 7 : AP-2 $\alpha$  and GATA-1 transcription factors bind to the first intron of *MUC5B*.**

Autoradiogram of the gel-shift assay in which radio-labelled oligonucleotides including the AP-2 (lanes 1-4) and GATA-1 (lanes 5-8) binding sites found in intron 1 were incubated with 5  $\mu$ g of Caco-2 nuclear extracts in the absence (lanes 2, 6) or presence (lanes 3, 4, 5, 8) of 1  $\mu$ l of specific antibody as indicated. Radio-labelled probe alone (lanes 1 and 7).

**Figure 8 : Sp1 and Sp3 effects on *MUC5B* promoter activity.**

Co-transfections were performed in enterocyte Caco-2 (white), mucus-secreting LS174T (black) and undifferentiated HT-29 STD (grey) colon cancer cell lines. 20  $\mu$ g of pGL3 deletion mutant 1596 or 1896 was co-transfected with 5  $\mu$ g of pCMV4 expression vector either empty (pCMV4) or carrying the coding sequence of Sp1 (pCMV4-Sp1) or Sp3 (pCMV4-Sp3) transcription factors. The results are means  $\pm$  S.D. and represent more than three different experiments in duplicate for each fragment.