Human mucin gene MUC5B is mapped clustered with MUC6, MUC2, and MUC5AC on chromosome 11p15.5. We report here the isolation of three overlapping genomic clones of human MUC5B spanning approximately 40 kilobases. We have determined their partial restriction maps and the intron-exon boundaries of the central region encoding a single open reading frame. This coding region has been completely sequenced. Its length is 10,713 base pairs, and it encodes a 3570-amino acid peptide. Nineteen subdomains have been individualized. Some subdomains show similarity to each other, creating larger composite repeat units that we have called super-repeats. Four super-repeats of 528 amino acid residues are thus observed within the central exon. Each comprises (i) a subdomain composed of 11 repeats of the irregular repeat of 29 amino acid residues, (ii) a unique conserved subdomain with no typical repeat, and (iii) a cysteine-rich subdomain. This latter subdomain has high sequence similarity to the cysteine-rich domains described in MUC2 and MUC5AC. Sequence data of these three genes, together with their clustered organization, lead us to suggest that they may be a part of a multigene family. The super-repeat present in MUC5B is the largest ever determined in mucin genes and the central exon of this gene is, by far, the largest reported for a vertebrate gene.

Mammalian respiratory, gastrointestinal, and reproductive tracts are protected by mucus secretions, of which the major components are the mucins. The mucins form a heterogeneous group of high molecular mass, polydisperse, highly glycosylated macromolecules. They are synthesized and secreted by specialized cells in the epithelium.

Considerable advances have been made over the past years toward our understanding of the structure and function of mucin glycoproteins. The isolation of mucin cDNA clones introduced a new approach to the structure of the mucins. Until now, at least eight human mucin genes have been identified (see Ref. 1 for review). The chromosomal localization of these genes has been established: four of them, MUC2, MUC5AC, MUC5B, and MUC6, are clustered on 11p15.5 between the HRAS and IGF2 genes, MUC1 is on 1q21–24, MUC3 on 7q22, MUC4 on 3q29, and MUC7 on chromosome 4q13–21. Recently, a cDNA called pAM1 has been cloned from a human tracheal library and localized on chromosome 12 (2). Three novel cDNAs have been reported: NP3a from a human nasal polyp library (3), L31 from a HT29-MTX cell line library (4), and HGM-1 from a human stomach library (5). Their sequences show that they correspond to some parts of the MUC5AC gene. More recently, a novel cDNA (pSM2–1) from human sublingual gland has been described (6). MUC1, which is developmentally regulated and aberrantly expressed by carcinomas, encodes a membrane-associated mucin-like glycoprotein. In contrast, the other described genes code for secreted mucins (1). Mucins present extended arrays of tandemly repeated sequences, producing a core protein core in potential O-glycosylation sites and having a high content of serine, threonine, proline, glycine, and alanine. The tandem repeat units vary in length from as few as 24 bp in MUC5AC (7) to 507 bp in MUC6 (8). The tandem repeat domain is flanked on either side by nonrepeat regions. In the MUC2 gene product, these tandem repeats are flanked by cysteine-rich subdomains of approximately 845 residues upstream and 700 residues downstream. Both cysteine-rich subdomains have sequences similar to the D-domains of human pro-von Willebrand factor (9, 10). Some parts of these D-domains are also found in NP3a (3) and in HGM-1 (5). Moreover, the MUC2 gene has upstream to the tandem repeat a region of imperfectly conserved repeats flanked by another type of cysteine-rich domain (11). This latter domain has also been described in HGM-1 (5), twice in MUC5AC (12), and also in its related cDNAs (3, 4).

Mucins are essential for the protective properties of the mucus (13, 14). In addition it is becoming apparent that an abnormal expression of the mucin genes occurs in various disease states and in conditions associated with a high risk of adenoma or carcinoma (15–17). Therefore, the study of the factors responsible for the regulation of mucin expression is of great interest. With this goal, it is necessary to acquire a detailed knowledge of the genomic structure of the mucin genes. The complete sequences of MUC1 (18), MUC2 (11, 9, 19), and MUC7 (20) cDNAs have been described. However, only partial cDNA sequences are published for the other mucin genes. The whole genomic structure is only known for MUC1 (21), although the partial genomic organization of MUC7 has
also been reported recently (22).
We have previously described four human tracheobronchial cDNA clones with degenerate 87-bp tandem repeats belonging to the human MUC5B gene (23). Our laboratory has focused considerable attention upon this gene which is expressed in mucous glands, gall bladder, and endocervix (24–26).

In this paper, we present the partial restriction map of the three overlapping genomic clones of the MUC5B gene spanning more than 40 kb and the complete sequence of its central exon which encompasses 10,713 bp and codes for a 3570-amino acid peptide. Nineteen subdomains are individualized. Some subdomains show similarity to each other, creating larger composite peptide. Nineteen subdomains are individualized. Some subdomains show similarity to each other, creating larger composite peptide. Nineteen subdomains are individualized.

**TABLE I**

| Bam HI | Bgl II | Eco RI | Hind III | Kpn I | Pst I | Xho I | Xho I |
|-------|--------|--------|----------|-------|-------|-------|-------|
| 1.6   | 18     | >30    | 25       | 8     | 1.4   | >30   | 25    |
| 1.1   | 2.2    | 1.25   | 3.7      | 2.1   | 0.95  | 1.6   |
| 0.4   | 1.6    | 0.65   |          |       |       |       |       |

The central exon of the human MUC5B gene, we screened a human placentall genomic DNA library in pWE15 cosmids provided by Stratagene using the JER57 probe. Two positive clones BEN1 and BEN2 were isolated and studied.

Restriction Mapping of Cosmids—The restriction mapping strategy of Wahl et al. (27) was slightly modified as follows. Cosmids were digested to completion and one part of this restriction site was prepared from leukocytes. It was digested with the following restriction endonucleases: Bam HI, Bgl II, Eco RI, Hind III, Kpn I, Pst I, Xho I, and Xho I. These were then subcloned into the pBluescript KS Kontakt primer), and 1.5 units of Taq DNA polymerase (Boehringer Mannheim). After overlying with 50 μl of mineral oil (Sigma), the mixture was denatured at 94°C for 3 min followed by 30 cycles at 94°C for 1 min, 71°C for 1 min, 72°C for 2 min. The elongation step was extended for an additional 10-min period. Secondary amplification was performed using 2.5 μl of the primary amplification product. The thermal cycling protocol used was the same as for the primary RACE amplification.

**DNA Extraction—** Total RNA was extracted from a human gall bladder using the guanidine isothiocyanate/CaCl method (28, 29).

**Reverse Transcription and Amplification—** A sample (0.5 μg) of human tracheal poly(A) RNA (Clontech) and a sample (1 μg) of total RNA extracted from human gall bladder were reverse-transcribed with the 5’-STRAND™ cDNA synthesis kit (Clontech) using random primers according to the manufacturer’s instructions.

The first strand cDNA (8 μl) and cosmid DNA (30 ng) were amplified by PCR with various primers: NAU112 (antisense) 5’-ACCGGCT- GGGCCGTGGCAGCGA-3’ (nt 3105–3129), NAU113 (sense) 5’-GAC- GACTACACGCCTGCCCAGTCCTAA-3’ (nt 1671–1687), NAU81 (sense) 5’-CAACTGCACTGCGACAGGGT-3’ (nt 694–716), NAU92 (antisense) 5’-GACGCTGGAGCACACGTGCA-3’ (nt 10601–10625), NAU128 (sense) 5’-CTGGTACCCTGTGTCCTCTGACTGTC-3’ (nt 10601–10625). They were sequenced using synthetic oligonucleotides corresponding to the T7 and T3 primers of the pK5 plastid, to the T7 and T3 primers of the pK5 plastid, and to 40 primers of the pGemT vector (Promega). Part of the sequence was determined by primer walking using primers specific to the MUC5B gene. Analyses of nucleic acid and protein sequence data were performed using PCGENE Software.

To perform DNA sequencing directly on cosmids (2 μg), we had to anneal at 37°C for 30 min and to use 5 pmol of primers instead of 0.5 pmol when sequencing DNA in plasmids using the Sequenase version 2.0. The nucleotide sequence reported in this paper has been submitted to the EMBL Data Bank with accession number X72496.

**RESULTS**

Southern Analysis—Human genomic DNA from leukocytes of 20 healthy unrelated volunteers was digested with Bam HI, Bgl II, Eco RI, Hind III, Kpn I, Pst I, Xba I, and Xho I. The sizes of the fragments obtained and hybridized with the JER57 probe are indicated in Table I. These results indicated the fragments of interest recognized by the JER57 probe. We isolated all these fragments from a phage or a cosmid genomic library and sequenced them to obtain the complete sequence recognized with the JER57 probe (see below).

**Isolation and Restriction Mapping of MUC5B Genomic DNA**

**Central Exon of the Human Mucin Gene MUC5B**

3169
Clones—The EMBL4 human genomic library has been screened with the JER57 probe, and one positive clone with an insert of approximately 12 kb called CEL5 was obtained. The JER57 probe hybridized with one BamHI-BamHI fragment of 1.1 kb (indicated in Figs. 1 and 2A) situated in the 5' part of this clone. This fragment was completely sequenced. Other screenings of the EMBL4 genomic library only gave the same clone. This led us to screen a human placenta genomic DNA clone. This fragment was completely sequenced. Other cosmid clones containing inserts of approximately 40 kb were isolated and sequenced. The partial restriction map of the three overlapping genomic clones—BEN1, BEN2, and CEL5—is indicated in Fig. 1. BEN1 contains a CpG island, since restriction sites obtained and called BEN1 and BEN2. The partial restriction map of the three clones (CEL5, BEN1, and BEN2) is indicated in Fig. 1. BEN1 contains a CpG island, since restriction sites such as BssHII and NotI are close together (30, 31). This island is located at 24 kb from the 5'-end of the insert and corresponds to the 18 CpG island on the macrocartography performed in the 11p15.5 region (32). BEN2 overlaps the 3' region of BEN1 on 16 kb and overlaps completely the CEL5 clone. We have found on these clones the restriction fragments corresponding to the fragments recognized with the JER57 probe and observed on Southern blots of genomic DNA (Fig. 1 and Table I).

Strategy for the Sequencing of the Cosmid Genomic Fragments Hybridizing with the JER57 Probe—We prepared, subcloned into pBluescript KS(+) and sequenced all the fragments indicated in Fig. 2A at the top. Various restriction fragments derived from these clones were also subcloned to determine the entire sequence. We distinguished the fragments obtained from BEN2 only (with an asterisk). One fragment BamHI-BamHI of 1.1 kb was obtained from BEN2 and from CEL5 (noted with an asterisk and ++ + + +). Since some fragments had the same lengths but were at various positions, we had to carefully isolate them starting from larger fragments unambiguously positioned, obtained for example only from BEN2, or for others only from BEN1. Due to an extremely high proportion of GC residues, many sequence problems were encountered. Subclones were thus sequenced several times before a reliable sequence was obtained.

Determination of Intron-Exon Boundaries on Each Side of the Central Exon—To obtain cDNAs upstream of the central tandem repeat region we used 5' RACE-PCR with the primers NAU58 and NAU57 chosen in a nonrepeat region in 5' of the central region. The amplification products were analyzed by agarose gel electrophoresis. Four ethidium bromide-stained bands ranging from 0.4 to 1 kb were obtained (data not shown) and cloned into pGEMT vector. Several transformants were isolated and sequenced. The nucleotide sequence showed that the 3' ends of all the clones sequenced overlap with the genomic sequence. The comparison of the sequence of the longest cDNA (985 bp) with the sequence of the fragment BamHI-KpnI of 1.35 kb (noted with a vertical arrow on the left in Fig. 2A) from the cosmid BEN1 indicates an intron of 468 bp (nucleotide sequence not shown), thus identifying the 5'-end of the central exon. Splice acceptor and donor sequences agree with the “GT-AG” rule (33).

To determine the 3'-end of the central exon we performed RT-PCR using two primers (NAU128 and NAU171, their positions are shown on Fig. 2A) on human RNAs from tracheobronchial tissue and from gall bladder. Then we compared the sequences obtained after subcloning into pMOSBlue vector with those determined on fragments of the cosmid located 3' of the central exon: fragment BamHI-SacII ~ 1.7 kb and PstI-SacII ~ 1.3 kb (noted with vertical arrows on the right in Fig. 2A). The presence of an intron at 27 bp downstream from the site PstI evidenced by comparing the sequences of the two cDNAs obtained with the genomic sequence. Splice acceptor and donor sequences agree with the GT-AG rule (33).

In order to confirm that the sequence was in the correct order, we performed restriction maps of overlapping fragments such as the 8-kb KpnI-KpnI fragment from BEN2 or the 5.5-kb HindIII-NotI 3'-end fragment from BEN1 (Fig. 1). To confirm that this sequence forms only one exon, additional PCR products were produced using various pairs of primers (see Fig. 2A). We compared the lengths and the sequences of the fragments obtained first by RT-PCR on gall bladder or tracheobronchial RNA and by PCR on genomic DNA (cosmids BEN1 or BEN2). Using NAU81 and NAU112 we obtained fragments of ~2500 bp, using NAU113 and NAU112 fragments of ~1400 bp, using NAU82 and NAU136 fragments of ~2100 bp. In each case, we obtained the same lengths and the same sequences with either cDNAs or genomic DNA as matrices. No intron was detected in this way. Thus the sequenced region consists of a single long open reading frame which extends over
10,713 bp (submitted to the EMBL Data Bank with accession number Z72496) coding for 3570 amino acid residues and leading to a polypeptide core with a calculated Mr of 370,000.

Comparison with the cDNAs Described Previously—The nucleotide sequences of the previously isolated cDNAs JUL10, JER28, and JUL7 (23) have been positioned within the central exon. JUL10 is at position 5272–6258. JUL7 overlaps JER28. They are respectively positioned at 8510–10143 and at 9172–9732. Some small differences in sequence were observed between the genomic sequence determined here and the cDNAs as determined previously by us (23). As we did not succeed to position the JER57 sequence as reported by us (23), we redetermined its nucleotide sequence (accession number X74955 with correction submitted to the EMBL Data Bank). The reason for this discrepancy between this newly determined sequence of JER57, which is at position 8361–10222, and our previous work is explained under “Discussion.”

These cDNAs were obtained by screening a human tracheobronchial tissue cDNA expression library using antibodies (23). Their deduced amino acid sequences allowed us to choose the appropriate open reading frame of the central exon continuous with that of their sequences.

Analysis of the Nucleotide Sequence and of the Deduced Amino Acid Sequence—Nineteen subdomains are individualized and are indicated in Fig. 2. Seven code for cysteine-rich subdomains called Cys1 (aa 6–112), Cys2 (aa 176–283), Cys3 (aa 457–565), Cys4 (aa 986–1093), Cys5 (aa 1515–1622), Cys6 (aa 2213–2320), and Cys7 (aa 2742–2849). Their amino acid sequences are displayed in Fig. 3. Their average amino acid composition is given in Table II (column Cys). These subdomains are rich in cysteine residues (9.3%). In Fig. 3 is also shown the similarity between these cysteine-rich subdomains and homologous domains found in other mucins such as human MUC5AC or related cDNAs (3, 5), mouse Muc5ac (34), pig gastric mucin (35), human MUC2 (9), and rat MUC2 homologue (36). In addition to the remarkable conservation of the cysteine residues, the conservation of numerous amino acid residues (bold boxes or boxes in Fig. 3) such as tryptophan, proline, arginine, and glycine is to be noted, especially between MUC5B and MUC5AC and its related cDNAs. There is one conserved potential O-glycosylation site in each of these cysteine-rich subdomains (boxed together with a tryptophan residue in the upper part of Fig. 3). No potential N-glycosylation site exists. The Cys4 to Cys7 subdomains of MUC5B mucin present a perfect sequence similarity except in three positions. The average amino acid composition of the seven cysteine-rich subdomains (see column Cys in Table II) is very similar to that of domain III in HGM-1 (5) (Fig. 3). It is noticeable that all but 2 of the cysteine residues of the central exon are in the cysteine-rich subdomains. 31 out of the 39 tyrosine residues present in the central exon are in the cysteine-rich subdomains.

The first two subdomains between Cys1-Cys2 and Cys2-Cys3, subdomains R01 (64 amino acid residues) and R02 (174 amino acid residues), respectively, are enriched with threonine, serine, proline, and alanine (Table II), but no typical repeats or even imperfect repeats can be discerned. Searching of the GenBank™ data base indicated that these sequences were not identical with any registered sequence. A certain similarity
with other mucins, especially MUC1 or MUC2, is due to the typical amino acid composition of mucins.

Five subdomains (RI to RV) composed of various numbers of 87-bp imperfect tandem repeats are encountered downstream of the regions coding for the cysteine-rich subdomains Cys3 to Cys7. The alignment of the corresponding nucleotide sequences is shown in Fig. 4A. Domains RI, RII, and RV are very similar. Each contains 11 imperfectly conserved repeats of 87 bp. The comparison is more obvious when looking at the deduced amino acid sequences (Fig. 4B). The subdomain RIII contains the same 11 irregular repeats of 29 amino acid residues (except RIII-7). Moreover, RII-1 to RII-5 are very similar to the first five repeats of subdomains RI, RII, and RV.

The subdomain RV is composed of 22 irregular repeats. RV-1 to RV-5 are also very similar to the first five repeats of the subdomains RI to RIV. RV-6 to RV-17 (except RV-7 and RV-9) are similar to the 11 imperfectly conserved repeats of the other R-subdomains of MUC5B with the insertion of the pentapeptide PTTTT in RV-14. Beneath the alignment is indicated the consensus sequence of this imperfectly conserved repeat and the percentages of the indicated amino acid residues. One serine residue has 100% conservation and 6 amino acid residues have more than 80% conservation. As far as all the five R-subdomains are concerned, 44 units have the 29-amino acid residue repeat out of the aligned 72 units. The others are composed of 28, 27, or 26 amino acid residues, although one contains 32

![Fig. 3. Amino acid sequence similarity between the cysteine-rich subdomains of human MUC5B mucin and other human or animal mucins. Sequences were deduced from nucleotide sequences and were aligned to give maximum identity with a minimal number of gaps. JUL32, JER47-1, and JER47-2 are parts of MUC5AC (12); domains III, V, and VII are from HGM-1 (5); NP3a (3); CRRI, and CRRII are from mouse Muc5ac (34); MUC2-1 and MUC2-2 (9); rat MUC2 homologue (36); PGM-2A is from pig (35). Conserved cysteine residues are shadowed. Bold boxed-in positions have 90–100% conservation, and boxed positions have from 70 to 90% conservation. Serine and threonine residues were boxed together.](image-url)
residues. The amino acid composition (Table II) shows an extremely high threonine content (37%) and high serine, proline, and alanine contents. There are numerous potential O-glycosylation sites. The sequence TXXP, which has been implicated as a major site for GalNAc addition, is found in most of the 72 repeats (5, 37). These regions are likely to be heavily glycosylated. Moreover, one potential repeat composed of the 11 imperfectly conserved repeats of 29 amino acid residues.

The alignment of these four super-repeats has been performed (not shown since the alignments of the various sequences existing in these super-repeats have already been shown in Figs. 3, 4B, and 5B). They present a striking similarity except in the 11 positions noticed before (in the R-end and the cysteine-rich subdomains) and in 54 positions in the first part of the super-repeat composed of the 11 imperfectly conserved repeats of 29 amino acid residues.

**DISCUSSION**

We report here the isolation of three overlapping genomic clones of the human MUC5B mucin, one from a phage library (CELS) and two from a cosmid library (BON1 and BON2) using the longest cDNA of MUC5B that we had previously isolated in the laboratory and called JER57 (23). MUC5B has been mapped clustered with MUC6, MUC2, and MUC5AC to chromosome 11p15.5. The partial restriction map of the three overlapping clones has been determined. We have also determined the intron-exon boundaries of the central region encoding a single open reading frame. This coding region has been completely sequenced. Its length is 10,713 bp, and it codes for 3570 amino acid residues. MUC5B is expressed in mucous glands of tracheobronchial tissue, submaxillary glands, gall bladder, and endocervix (24–26). RT-PCR experiments carried out on total RNA from human gall bladder or on human tracheal poly(A[-]) RNA unambiguously demonstrated that this open reading frame consists of a single exon and does not contain any intron. It was not possible to amplify the complete length of this open reading frame by RT-PCR using the two primers, NAU81 and NAU92, that are on both ends of the central exon. However, by performing different amplifications using more closely situated primers, in particular in the MUC5B specific part of the cysteine-rich domains, we were able to produce fragments exhibiting the same lengths and the same sequences as those in the cosmid clones. No difference was found in the common sequences determined on the genomic clones from the two libraries, even though they came from two unrelated individuals. This gene differs from MUC1 and MUC2 for which variable numbers of random repeats polynucleotides have been demonstrated (38, 11). Since the repeat unit of 87 bp is imperfectly

---

**TABLE II**

| Amino acid | R01 | R02 | Cys | R | R-end | R03 |
|------------|-----|-----|-----|---|-------|-----|
|            | Nb  | %   | Nb  | % | Nb    | %   |
| Ala        | 7   | 11.1| 16  | 9.2| 38    | 5.0 |
| Arg        | 1   | 1.6 | 8   | 4.6| 44    | 5.8 |
| Asn        | 0   | 0.0 | 0   | 0.0| 28    | 3.7 |
| Asp        | 0   | 0.0 | 0   | 0.0| 37    | 4.9 |
| Cys        | 0   | 0.0 | 0   | 0.0| 70    | 9.3 |
| Glu        | 3   | 4.8 | 6   | 3.5| 43    | 5.7 |
| Gly        | 2   | 3.2 | 8   | 4.6| 58    | 7.7 |
| His        | 4   | 6.3 | 12  | 6.9| 69    | 9.7 |
| Ile        | 1   | 1.6 | 0   | 0.0| 15    | 2.0 |
| Lys        | 1   | 1.6 | 0   | 0.0| 17    | 2.2 |
| Met        | 0   | 0.0 | 2   | 1.2| 17    | 2.2 |
| Phe        | 0   | 0.0 | 1   | 0.6| 31    | 4.1 |
| Pro        | 10  | 15.9| 24  | 13.9| 55    | 7.3 |
| Ser        | 11  | 17.5| 23  | 13.3| 41    | 5.4 |
| Thr        | 15  | 23.8| 51  | 29.5| 32    | 4.2 |
| Trp        | 1   | 1.6 | 0   | 0.0| 17    | 2.2 |
| Tyr        | 0   | 0.0 | 1   | 0.6| 31    | 4.1 |
| Val        | 4   | 6.3 | 3   | 1.7| 50    | 6.6 |
| Total      | 63  | 100 | 173 | 100| 756   | 100 |

The number of residues (Nb) and their percentage (%) are indicated for each subdomain type. The first five amino acid residues of the central exon are not included. Results for amino acid residues of interest (Cys, Ser, Thr) are in bold type.
conserved, we had to sequence the whole central exon, while in MUC2 the authors were able to extrapolate the sequence repeated up to 100 times (9, 11).

The length of the central exon is much larger than that of MUC1, which varies from 3.5 to 6.2 kb, depending on the number of repeats (38). MUC7, a nonforming-gel mucin, possesses a central exon of 2.2 kb (22). As far as MUC2 is concerned, the intron/exon distribution is not as yet known. Toribara et al. (11) have indicated that the 5' region of the GMUC clone isolated in their laboratory forms, together with

FIG. 4. Nucleotide sequences (A) and deduced amino acid sequences (B) of the R-subdomains. The residues of serine and threonine are shadowed in B.
the tandem repeat array, a single large exon. In the most common genotype for MUC2, this region extends over 8,700 bp. No information is available about the size of the exonic repeat region of the other mucin genes. A survey on intron and exon lengths has indicated that in vertebrates few exons are over 800 bp (39). Most vertebrate genes are composed of long introns.
and short exons. Composed of 10,713 bp, the central exon of MUC5B is even much larger than the 7572-bp exon of the gene for lipoprotein ApoB considered to be the largest one in vertebrates (40). The corresponding polypeptide coded by the central exon of MUC5B has a calculated Mr of 370,000.

Nineteen subdomains have been individualized. Seven subdomains of 108 amino acid residues, called Cys1 to Cys7, contain 10 cysteine residues and show a very similar organization to that observed twice in human (11) and rat (36) MUC2, at least twice in human (12), and mouse (34) MUC5AC as well as in other mucin cDNAs, which can now be identified as representing parts of the MUC5AC gene, i.e., NP3a (3), HGM-1 (5), FIG. 5. Nucleotide sequences (A) and deduced amino acid sequences (B) of the R-end subdomains. In A, identical nucleotide sequences are shaded, and a consensus sequence is given: R is A or G, Y is T or C, M is A or C, and S is C or G. The residues of serine and threonine are shadowed in B.
and L31 (4). The structure of the cysteine-rich subdomains has been conserved over a long evolutionary time scale, and the evolutionary constraint has likely been maintained because of crucial disulfide bonds. Thus, this typical cysteine-rich domain seems to be a feature of at least three of the four human mucin genes located on 11p15.5, little information is available as to whether there is a similar domain in MUC6 (8). It is more and more tempting to speculate, as have Toribara et al. (8), that at least MUC2, MUC5AC, and MUC5B genes on chromosome 11p15.5 are part of a multigene family whose members code for secreted mucins. The conservation of numerous amino acid residues in addition to the cysteine residues suggests an important role for these domains. It is noticeable that the MUC5B and MUC5AC genes show greater similarity with each other than with the MUC2 genes (from human and rat). Deletions of amino acid residues occur at the same positions in MUC5B and MUC5AC, while some observed deletions are characteristics of the MUC2 type. Moreover, MUC5B and MUC5AC are very close together in the cluster (32). Overall, these observations suggest that MUC5AC and MUC5B may have the same ancestral gene.

It is noteworthy that the mouse mucin gene homologous to human MUC5AC is mapped to a site on mouse chromosome 7 homologous to the location of the human secretory mucin gene cluster on human chromosome 11p15.5 (34). We do not as yet know if mouse possesses both Muc5ac and Muc5b. Perhaps only one gene exists.

The presence of multiple cysteine residues strongly supports the idea that the MUC5B gene codes for a secreted mucin as disulfide bonding is necessary for the formation of a mucous gel. Moreover, we can postulate that interactions may occur via cysteine-rich domains between several mucin molecules producing a complicated network. This can explain the observations made using electron microscopy of bronchial mucins which have been shown to form complex entangled structures (41) or “bush-like” aggregates (42). The cysteine-rich domains may play a highly conserved role such as the packaging of the mucins or in mediating interactions with numerous other proteins such as, for example, the association with bilirubin in the gall bladder matrix (43). Further studies will be required to more adequately clarify the role of these domains.

An examination of the newly determined sequence of JER57 cDNA led us to conclude that the changes in the reading frame emphasized in our previous work (23) were in fact due to errors in sequence determination because of difficulties encountered in determining the sequence of such imperfectly conserved repeat domains. This is a problem that we have now overcome. We will now call the 87-bp repeat “imperfectly conserved” or “irregular” rather than “degenerate” repeat as designated previously. The cDNAs previously isolated (23) essentially consist of parts of the R-subdomain. In fact, the repeat of 29 amino acid residues of the R-subdomains appears to be only one of the components of the super-repeat. This super-repeat is the largest ever determined in mucin genes. The largest described until now was the 507-bp repeat unit of MUC6 (8). The super-repeat in MUC5B is more than three times as long as the MUC6 repeat. Its particularity is that it is made of three various subdomains. A variable number of this super-repeat has not been as yet observed. An interesting aspect of the deduced polypeptide sequence is the alternating arrangement of the three types of subdomains. The four R-end subdomains show a striking similarity to each other. Their role is likely to be crucial.

The conservation of the potential N-glycosylation sites within the repeat may be important for the proper maturation of MUC5B providing the positions at which N-glycosylation occurs. Recently, Denny et al. (44) reported that cloning of mucin apoproteins showed that the consensus sequence(s) for N-glycosylation is usually found outside of the repeat domain. For MUC5B, five out of the seven potential N-glycosylation sites are within the repeat region. The subdomains R01, R02, and R03 display amino acid compositions typical of mucins but do not contain any repeat.

The MUC5B gene is clustered in 11p15.5 with MUC2, MUC6, and MUC5AC (8, 32). Although mucins are characterized by having tandem repeat regions containing significant amounts of threonine, serine, alanine, and proline, the individual repeat units for each of these genes are very different in length and amino acid sequence. Moreover, the number of the cysteine-rich subdomains emphasized here, which are different from the D-domains of human pro-von Willebrand factor (10), considerably differs between MUC2 and MUC5B. The role of these two types of secreted mucins may be highly specialized as suggested by in situ hybridization experiments which show a different expression pattern for each gene (24). In fact, we will now be able to design new oligonucleotides and produce fusion proteins and antibodies in order to reinvestigate the expression of the MUC5B gene using in situ hybridization and immunohistochemistry.

Experiments performed to elucidate the entire genomic organization of MUC5B are now progressing as is the study of the regulatory regions. Whether the regulatory elements of MUC5B function coordinately with the regulatory elements of adjacent mucin genes in the cluster is an important question to answer in the future.

Acknowledgments—We are indebted to Dr. A. T. Norden for help in improving the style of the paper. We thank Pascal Mathon and Daniele Petitprez for technical assistance.

REFERENCES

1. Gendler, S. J., and Speicer, A. P. (1995) Annu. Rev. Physiol. 57, 607–634
2. Shankar, V., Gilmore, M. S., Elkins, R. C., and Schulder, G. P. (1994) Biochem. J. 300, 295–298
3. Meerzaman, D., Charles, P., Daskal, E., Polymeroopoulos, M. H., Martin, B. M., and Rose, M. C. (1994) J. Biol. Chem. 269, 12992–12999
4. 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
5. Klomp, L. W. J., Van Bens, L., and Straus, G. J. (1995) Biochem. J. 308, 831–838
6. Troxler, F. R., Offner, G. D., Zhang, F., Iontcheva, I., and Oppenheim, F. E. (1995) Biochem. Biophys. Res. Commun. 217, 1112–1119
7. Aubert, J. P., Pochet, N., Ougrin, M., Dutreque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D., and Degand, P. (1991) Am. J. Respir. Cell Mol. Biol. 5, 178–185
8. Toribara, N. W., Robertson, A. M., Ho, J. W., B., Kuo, W.-L., Gum, E., Hicks, J. W., 208, 20563–20569
9. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R., and Kim, Y. S. (1993) J. Biol. Chem. 268, 5879–5885
10. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R., and Kim, Y. S. (1995) J. Biol. Chem. 264, 19134–19127
11. Toribara, N. W., Gum, J. R., Culhane, M., Lagace, R. E., Hicks, J. W., Petersen, G. M., and Kim, Y. S. (1991) J. Biol. Chem. 266, 1005–1013
12. Goyonnet-Dupere, V., Audie, J. P., Debaillie, L., Laine, A., Biwsine, M. P., Zouitina-Galie`gue, S., Pigny, P., Degand, P., Aubert, J. P., and Perchet, N. (1995) Biochem. J. 305, 211–219
13. Neutra, M. R., and Forstner, J. F. (1987) in Physiology of the Gastrointestinal Tract (Johnson, L. R., ed) Raven Press, New York, pp 325–336
14. Van Klinken, B. J. W., Dekker, J., Muller, H. A., and Enerterman, A. W. C. (1995) Am. J. Physiol. 269, G613–G627
15. Ho, J. L., and Kim, Y. S. (1995) Int. J. Oncol. 7, 913–926
16. Lesuffleur, T., Zweibaum, A., and Real, F. X. (1994) Crit. Rev. Oncol. Hematol. 17, 153–180
17. Biwsine, M. P., Janin, A., Masumoury, V., Audie, J. P., Debaillie, L., Laine, A., Biwsine, M. P., Zouitina-Galie`gue, S., Pigny, P., Degand, P., Aubert, J. P., and Perchet, N. (1996) Gastroenterology 110, 84–91
18. Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rotherold, J., and Burchell, J. (1996) J. Biol. Chem. 261, 12820–12827
19. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) J. Biol. Chem. 269, 2440–2446
20. Bobek, L. A., Tsa, H., Biesbroek, A., and Levine, M. J. (1993) J. Biol. Chem. 268, 20563–20569
21. Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990) Biochem. Biophys. Res. Commun. 173, 1019–1029
22. Bobek, L. A., Liu, J., Sait, S. N., Shens, T. B., Bobek, Y. A., and Levine, M.
Central Exon of the Human Mucin Gene MUC5B

J. (1996) Genomics 31, 277–282

23. Dufosse, J., Porchet, N., Audie, J. P., Guyonnet-Dupérat, V., Laine, A., Van-Steuningen, I., Marrakchi, S., Degand, P., and Aubert, J. P. (1993) Biochem. J. 293, 329–337

24. Audie, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B., and Aubert, J. P. (1993) J. Histochem. Cytochem. 41, 1479–1485

25. Audie, J. P., Tétaert, D., Pigny, P., Buisine, M. P., Janin, A., Aubert, J. P., Porchet, N., and Boersma, A. (1995) Hum. Reprod. 10, 98–102

26. Campion, J. P., Porchet, N., Aubert, J. P., L’Helgoualc’h, A., and Clément, B. (1995) Hepatology 21, 223–231

27. Wahl, G. M., Lewis, K. A., Ruiz, J. C., Rothenberg, B., Zhao, J., and Evans, G. A. (1987) Proc. Natl. Acad. Sci. U. S. A. 84, 2160–2164

28. Glisin, V., Orkvenjakov, R., and Byus, C. (1974) Biochemistry 13, 2633–2637

29. Crépin, M., Porchet, N., Aubert, J. P., and Degand, P. (1990) Biorheology 27, 471–484

30. Brown, W. R. A., and Bird, A. P. (1986) Nature 322, 477–481

31. Bickmore, W. A., and Bird, A. P. (1992) Methods Enzymol. 216, 224–244

32. Guyonnet-Dupérat, V. (1993) Study of the Human Mucin Genes Located in 11p15. Ph.D. thesis, University of Lille, France

33. Jacob, M., and Gallinaro, H. (1989) Nucleic Acids Res. 17, 2159–2180

34. Shekels, L. L., Lyftogt, C., Kieliszewski, M., Filie, J. D., Kozak, C. A., and Ho, S. B. (1995) Biochem. J. 311, 775–785

35. Turner, B. S., Bhaskar, K. R., Hadzopoulou-Cladaras, M., Specian, R. D., and Lamont, J. T. (1995) Biochem. J. 308, 89–96

36. Ohmori, H., Dohrman, A. F., Gallup, M., Tauda, T., Kai, H., Gum, J. R., Jr., Kim, Y. S., and Basbaum, C. B. (1994) J. Biol. Chem. 269, 17833–17840

37. Elhammer, A. P., Poorman, R. A., Brown, E., Maggiorni, L. L., Hoogerheide, J. G., and Keedy, F. J. (1993) J. Biol. Chem. 268, 10029–10038

38. Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E.-N., and Wilson, D. (1996) J. Biol. Chem. 265, 15286–15293

39. Hawkins, J. D. (1988) Nucleic Acids Res. 16, 9893–9908

40. Blackhart, B. D., Ludwig, E. M., Pierotti, V. R., Caiati, L., Onasch, M. A., Wallis, S. M., Powell, L., Pease, R., Knott, T. J., Chu, M.-L., Mahley, R. W., Scott, J., McCarthy, B. J., and Levy-Wilson, B. (1986) J. Biol. Chem. 261, 15364–15367

41. Sheehan, J. K., Oates, K., and Carlstedt, I. (1986) Biochem. J. 239, 147–153

42. Slayter, H. S., Lamblin, G., Le Treut, A., Galabert, C., Houdret, N., Degand, P., and Roussel, P. (1984) Eur. J. Biochem. 142, 209–216

43. Smith, B. J., and LaMont, J. T. (1985) J. Clin. Invest. 76, 439–445

44. Denny, P. C., Mirels, L., and Denny, P. A. (1996) Glycobiology 6, 43–50