BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **251,** 350–355 (1998)
ARTICLE NO. RC989469

The Amino-Terminal Sequence of MUC5B Contains Conserved Multifunctional D Domains: Implications for Tissue-Specific Mucin Functions

Gwynneth D. Offner,*[,1] David P. Nunes,* Andrew C. Keates,† Nezam H. Afdhal,* and Robert F. Troxler‡

*Departments of *Medicine and ‡Biochemistry, Boston University Medical Center, Boston, Massachusetts; and †Section of Gastroenterology, Department of Medicine, Beth Israel Hospital and Harvard Medical School, Boston, Massachusetts*

Received August 25, 1998

The MUC5B mucin gene product is expressed in a wide variety of secretory epithelia including the gallbladder, salivary glands, trachea, and colon. Previous studies by us and others have described the C-terminal region as well as the central tandem repeating domain of this mucin. In an effort to understand the functional role of MUC5B in diverse human tissues, the sequence encoding the N-terminal region of this mucin was determined from the sequences of exons in three overlapping genomic clones. Primer extension mapped the site of transcription initiation 25 bp downstream from a putative TATA box element. The N-terminal region of MUC5B contained 1321 amino acids organized into a signal peptide, a short serine-threonine rich region, and three von Willebrand factor-like D domains. Comparison of this sequence with the N-terminal sequences of MUC2 and MUC5AC revealed a much higher degree of identity (46–59%) than that observed in the C-terminal regions of these mucins (33%). The N-terminal sequence of MUC5B also contains a number of sequence motifs common to several groups of extracellular ligand binding and adhesion proteins not previously recognized in mammalian gel-forming mucins. The N-terminal D domains in MUC5B are likely to have important roles in both mucin assembly and in the protective functions of the secreted mucin. © 1998 Academic Press

Mucins are large multimeric glycoproteins which are secreted by epithelial surfaces lining the gastrointestinal, respiratory and genitourinary tracts. At least nine genes for distinct human mucins have been identified to date: MUC1-4, MUC5AC, MUC5B, and MUC6-8 (1,2). Mucin gene products are expressed in tissue specific, though overlapping patterns and are thought to protect the underlying epithelial surfaces from microbial, physical and chemical assault. MUC5B has one of the widest tissue distributions of known mucins. It was first identified in a human trachea cDNA library (3). Subsequently, we showed that it represents the major human gallbladder mucin (4) as well as the major mucin component of the high molecular weight salivary mucin MG1 (5-7). MUC5B has also been identified in tissues of the female reproductive tract (8) and the colon (2,3).

MUC5B is one of the four known mucin genes clustered on human chromosome 11p15.5 along with MUC2, MUC5AC, and MUC6 (9). The first of these mucins for which the complete structural organization was determined was MUC2 (10, 11). This mucin was comprised of a central domain containing O-glycosylated tandem repeating sequences which was flanked at the N- and C-termini by cysteine-rich sequences with a high degree of similarity to D domains in human von Willebrand factor (12). Subsequently, the C-terminal sequences of MUC5AC (13,14), MUC5B (4, 6) and MUC6 (15) and the N-terminal sequence of MUC5AC (16-18) were determined. With the exception of MUC6, these data revealed that the 11p15 mucins share a conserved structure with MUC2 and von Willebrand factor.

Given the wide tissue distribution of MUC5B, it was of considerable interest to analyze the N-terminal sequence of this mucin in order to gain insights into its function in diverse biological environments. In this report, we describe a series of three overlapping genomic and P1 clones from which the complete nucleotide sequence of the N-terminal region of MUC5B was determined. This sequence, organized into three D do-

[1] To whom correspondence should be addressed at: Section of Gastroenterology, Boston Medical Center Evans 201, 88 E. Newton St., Boston, MA 02118. Fax: 617 638-8224. E-mail: goffner@acs.bu.edu.

The sequence presented in this paper has been deposited in GenBank with the Accession Number AF086604.

0006-291X/98 $25.00

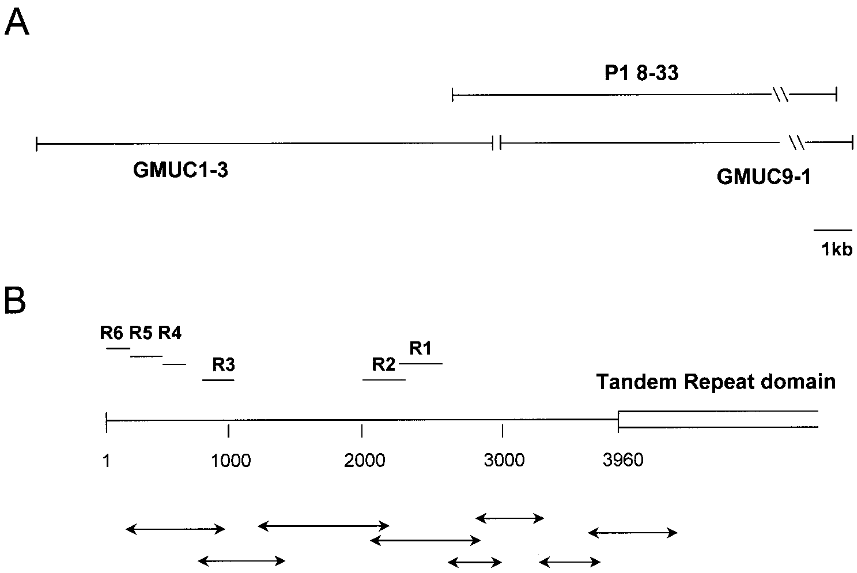


FIG. 1. Strategy used to determine the N-terminal sequence of MUC5B. (A) Schematic representation of the genomic (GMUC1-3 and GMUC9-1) and P1 (P1 8-33) clones which were sequenced to identify MUC5B exons. (B) Schematic representation of the PCR products (shown as arrows) and 5′ RACE products (R1-R6) used to confirm the exon sequences identified in (A). Only those PCR products required to obtain the complete sequence are shown.

mains, was compared to the corresponding sequences of MUC5AC and MUC2 and a strikingly high degree of sequence identity was observed. This similarity suggests that the N-terminal domain of MUC5B may be critical for both the structural assembly and function of this mucin.

MATERIALS AND METHODS

Human genomic and P1 library screening and DNA sequencing. A human genomic DNA library in lambda Dash II was purchased from Stratagene (La Jolla, CA) and human P1 library filters were purchased from Genome Systems (St. Louis, MO). The human genomic library was first screened with a random primer labeled probe encoding MUC5B tandem repeats (clone hGBM 4-1; 4) to yield clone GMUC 9-1. The same genomic library was subsequently screened with a PCR-derived fragment from P1 clone 8-33 (see below) to provide clone GMUC 1-3. Hybridization and wash conditions used for screening the genomic library were as described previously (4). P1 library filters were screened with a 1.5 kb fragment obtained by PCR from the 5′ end of genomic clone GMUC 9-1 using hybridization and high stringency was conditions were as described by the supplier. DNA was isolated from large scale liquid lysate cultures of both the genomic and P1 clones and sequence analysis was performed directly on these DNA samples using vector or sequence-specific oligonucleotide primers with an Applied Biosystems 373A automated instrument. In other experiments restriction fragments from these DNA samples were subcloned into pBluescript and sequenced as described above.

RT-PCR and 5′ RACE. RNA was isolated (19) from human gallbladder tissue obtained at cholecystectomy or from a human gallbladder epithelial cell line (20). For RT-PCR reactions, RNA (1 μg) was reverse transcribed in reactions containing random hexamer primers (50 pmol) or sequence specific primers (10 pmol) and the resulting cDNA used in standard PCR reactions. All 5′ RACE procedures were performed using RNA isolated from the gallbladder epithelial cell line with a kit from Gibco-BRL (Gaithersburg, MD). RT-PCR and RACE products were cloned into pCR II (Invitrogen, Carlsbad, CA) and sequenced as described above.

Primer extension. Primer extension reactions were carried out to verify the transcription start site identified in the sequence of GMUC1-3 (see below). RNA from the human gallbladder epithelial cell line (50 μg) was reverse transcribed using an AMV reverse transcriptase system (Promega, Madison, WI) using end labeled primer SP7 (underlined in Figure 3). Extension products were treated with RNase A, phenol-chloroform extracted and ethanol precipitated. The pellet was resuspended in loading dye, heated at 90°C for 10 min and electrophoresed on a 6% acrylamide-7M urea gel next to an M13 sequencing ladder.

RESULTS

Determination of MUC5B N-terminal sequence. The overall strategy used to obtain the complete nucleotide sequence of the 5′ region of MUC5B is shown in Figure 1. This strategy involved sequencing two genomic clones, pGMUC9-1 and pGMUC1-3 and P1 clone P18-33 by “walking” from both ends, or from a defined point (see below) using specific oligonucleotide primers. The sequence data obtained were analyzed using the BCM Gene Finder program to predict intron/exon boundaries and oligonucleotide primers corresponding to predicted exon sequences were synthesized. These primers were used in RT-PCR reactions to generate cDNA fragments containing the coding region of the N-terminal region of MUC5B. A series of 5′ RACE reactions was carried out concomitantly to confirm the positions of exon/intron boundaries and to identify the 5′ end of the MUC5B mRNA.

The first genomic clone isolated, GMUC9-1 contained sequences coding for MUC5B beginning with residue 1099 (Figure 2) and extending into the sequence of the central 10.7 kb exon of MUC5B described by Desseyn et al (21). A 1.5 kb PCR product derived from the 5′ end of this clone was used to screen a set of P1 library filters. Clone P1 8-33 was sequenced from the 5′ end using the SP6 vector primer and from a the 3′ end using a primer which was contained in GMUC9-1. This P1 clone was found to contain sequences coding for MUC5B beginning with residue 960 (Figure 2) and overlapping with GMUC 9-1. A 0.7 kb PCR product from the 5′ end of the P1 clone was used to re-screen the human genomic library and clone GMUC1-3 was isolated. This clone was sequenced from both ends by walking with T3, T7 and specific oligonucleotide primers and was found to contain sequences coding for residues 1-1098 of MUC5B.

Analysis of the deduced N-terminal sequence of MUC5B. The nucleotide sequence of MUC5B upstream of the central exon described by Desseyn et al. (21) contains 3963 nucleotides and codes for 1321 amino acids (Figure 2). The sequence begins with a variant Kozak (22) consensus translation start sequence (GGATGG) followed by a 25 amino acid sequence resembling a typical signal peptide and a 50 residue sequence enriched with serine, threonine and proline. Analysis of the latter sequence using the Net-O-Glyc program reveals several potential O-glycosylation sites. This mucin-like sequence is followed by a cysteine-rich sequence which contains 3 domains similar to the D domains of human pre pro-von Willebrand factor (12). The D1 (residue 76-422), D2 (residue 423-778) and D3 (residue 893-1249) domains contain 11 potential N-glycosylation sites and structural analyses predict that these domains have the potential to form globular ordered structure. Following the D3 domain is a 71 amino acid cysteine-rich sequence which connects to the central tandem repeat domain.

Comparison of the N-terminal sequence of MUC5B with the corresponding sequences of MUC5AC (16–18) and MUC2 (10–11) reveals a very high degree of sequence identity (Figure 2). MUC5B displays 58.6% and 45.7% identity with MUC5AC and MUC2, respectively whereas MUC5AC and MUC2 are 46.8% identical. This is significantly higher than that found in the C-terminal regions of these mucins which are approximately 33.% identical. Most notably conserved in the N-terminal sequences are the positions of cysteine residues where of the 124 cysteines present in MUC5B, 105 are conserved in all three mucins. Further analysis of the N-terminal sequences of MUC5B, MUC5AC and MUC2 revealed that the sequence identity between them was greatest in the D3 domain (45%) with 35% and 41% similarity in the D1 and D2 domains, respectively.

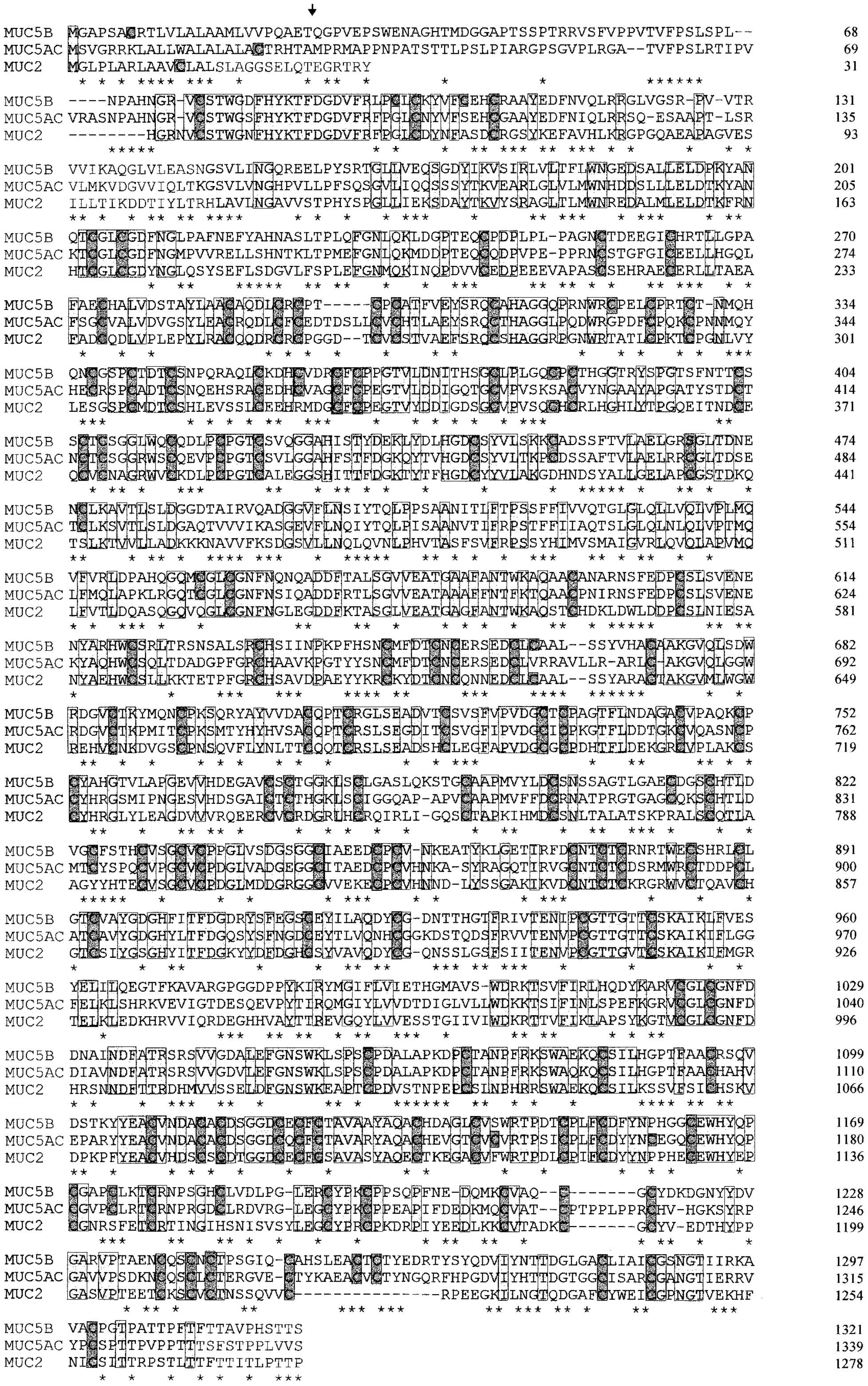
An analysis of the deduced sequence of the N-terminal region of MUC5B using the IDENTIFY program revealed that the D3 domain contained a consensus motif for C-type lectin domain proteins as well as a motif common to selectin complement binding repeats. In addition, a search of GenBank revealed that all 3 D domains contained a sequence with similarity to a block of 4 LDL receptor Class A repeats present in the LDL receptor family proteins megalin and the a_2 macroglobulin receptor. In addition, D domain sequences were also identified which were similar to the cysteine-rich EGF repeats in the extracellular proteins laminin and tenascin and the calcium binding EGF repeats in human fibrillin and in the Drosophila neurogenic notch protein. The 71 amino acid linker region following the D3 domain also displayed significant similarity to the EGF repeats in the proteoglycan core protein perlecan.

Primer extension. Primer extension reactions were carried out with primer SP7 (underlined in Figure 3) and RNA isolated from a human gallbladder epithelial cell line. A major 101 base extension product was observed (Figure 3A) which identifies the transcription start site as an "A" 25 bases downstream from the TATA box-like element identified in the sequence of GMUC1-3 and 57 bases upstream from the ATG coding for the first residue of the predicted signal peptide (Figure 3B).

DISCUSSION

In this report, we describe the sequence of the N-terminal region of MUC5B and show that the N-terminal regions of the 11p15 mucins MUC5B, MUC5AC and MUC2 display a much greater degree of sequence identity than that found in the C-terminal regions of these mucins. This striking finding suggested that structural features in this region required for assembly or function have been conserved during evolution of this mucin gene family.

The assembly of human pre pro-von Willebrand factor multimers occurs by interchain disulfide linkages between D3 domains (23). This process is self-catalyzed and is thought to involve the sequence CGLCG which is similar to a sequence at the active site of protein-disulfide isomerase (24). Recent studies have shown that assembly of porcine submaxillary mucin multimers also involves intermolecular disulfide linkages between N-terminal D domains, although it is not known which of the D domains participates in this process (25). The D1, D2 and D3 domains of MUC5B and MUC5AC each contain the CGLCG sequence whereas this sequence is present in only the D1 and D3 domains of MUC2. Comparison of the individual D domains in



```
                               ↓
MUC5B  MGAPSACRTLVLALAAMLVVPQAETQGPVEPSWENAGHTMDGGAPTSSPTRRVSFVPPVTVFPSLSPL--   68
MUC5AC MSVGRRKLALLWALALALACTRHTAMPRMAPPNPATSTTLPSLPIARGPSGVPLRGA--TVFPSLRTIPV   69
MUC2   MGLPLARLAAVCLALSLAGGSELQTEGRTRY                                          31
        * * * **** ***  *      * **  *       *         *          ******

MUC5B  ----NPAHNGR-VCSTWGDFHYKTFDGDVFRLPCICKYVFCEHCRAAYEDFNVQLRRGLVGSR-PV-VTR   131
MUC5AC VRASNPAHNGR-VCSTWGSFHYKTFDGDVFRFPGLCNYVFSEHCGAAYEDFNIQLRRSQ-ESAAPT-LSR   135
MUC2   --------HGRNVCSTWGNFHYKTFDGDVFRFPGLCDYNFASDCRGSYKEFAVHLKRGPGQAEAPAGVES   93
           *****                      * *    *  ** *** ** *** * *     *   * *

MUC5B  VVIKAQGLVLEASNGSVLINGQREELPYSRTGLLVEQSGDYIKVSIRLVLTFLWNGEDSALLELDPKYAN   201
MUC5AC VLMKVDGVVIQLTKGSVLVNGHPVLLPFSQSGVLIQQSSSYTKVEARLGLVLMWNHDDSLLLELDTKYAN   205
MUC2   ILLTIKDDTIYLTRHLAVLNGAVVSTPHYSPGLLIEKSDAYTKVYSRAGLTLMWNREDALMLELDTKFRN   163
       ** *  * ** ** ****     * *  *   * ***    *     ** ***   * ***    * **

MUC5B  QTCGLCGDFNGLPAFNEFYAHNASLTPLQFGNLQKLDGPTEQCPDPLPL-PAGNCTDEEGICHRTLLGPA   270
MUC5AC KTCGLCGDFNGMPVVRELLSHNTKLTPMEFGNLQKMDDPTEQCQDPVPE-PPRNCSTGFGICEELLHGQL   274
MUC2   HTCGLCGDYNGLQSYSEFLSDGVLFSPLEFGNMQKINQPDVVCEDPEEEVAPASCSEHRAECERLLTAEA   233
               *  **    *****  ** **   *   *  ***     ** ** * *   ** ***  * *

MUC5B  FAECHALVDSTAYLAACAQDLCRCPT-----CPCATFVEYSRQCAHAGGQPRNWRCPELCPRTCT-NMQH   334
MUC5AC FSGCVALVDVGSYLEACRQDLCFCEDTDSLLCVCHTLAEYSRQCTHAGGLPQDWRGPDFCPQKCPNNMQY   344
MUC2   FADCQDLVPLEPYLRACQQDRCRCPGGD--TCVCSTVAEFSRQCSHAGGRPGNWRTATLCPKTCPGNLVY   301
        *   *  *           * * *  *    *    * *            *   * *   * *  ***

MUC5B  QNCGSPCTDTCSNPQRAQLCKDHCVDRCFCPPGTVLDNITHSGCLPLGQCPCTHGGTRYSPGTSFNTTCS   404
MUC5AC HECRSPCADTCSNQEHSRACEDHCVAGCFCPEGTVLDDIGQTGCVPVSKSACVYNGAAYAPGATYSTDCT   414
MUC2   LESGSPCMDTCSHLEVSSLCEEHRMDGCFCPEGTVYDDIGDSGCVPVSQCHCRLHGHLYTPGQEITNDCE   371
        ***        * * * * **  ***    *   * * * *  * ***                 **

MUC5B  SCTCSGGLWQCQDLPCPGTCSVQGGAHISTYDEKLYDLHGDCSYVLSKKCADSSFTVLAELGRSGLTDNE   474
MUC5AC NCTCSGGRWSCQEVPCPGTCSVLGGAHFSTFDGKQYTVHGDCSYVLTKPCDSSAFTVLAELRRCGLTDSE   484
MUC2   QCVCNAGRWVCKDLPCPGTCALEGGSHITTFDGKTYTFHGDCYYVLAKGDHNDSYALLGELAPCGSTDKQ   441
         * ** *   ***      **   * ** * *   *     *      *  ***** *   ** *   *

MUC5B  NCLKAVTLSLDGGDTAIRVQADGGVFLNSIYTQLPPSAANITLFTPSSFFIVVQTGLGLQLLVQLVPLMQ   544
MUC5AC TCLKSVTLSLDGAQTVVVIKASGEVFLNQIYTQLPISAANVTIFRPSTFFIIAQTSLGLQLNLQLVPTMQ   554
MUC2   TSLKTVVLLADKKKNAVVFKSDGSVLLNQLQVNLPHVTASFSVFRPSSYHIMVSMAIGVRLQVQLAPVMQ   511
       **    * ** *  **** ***   *  *****   ** * *  *  ***   ** * **  *  *

MUC5B  VFVRLDPAHQGQMCGLCGNFNQNQADDFTALSGVVEATGAAFANTWKAQAACANARNSFEDPCSLSVENE   614
MUC5AC LFMQLAPKLRGQTCGLCGNFNSIQADDFRTLSGVVEATAAAFFNTFKTQAACPNIRNSFEDPCSLSVENE   624
MUC2   LFVTLDQASQGQVQGLCGNFNGLEGDDFKTASGLVEATGAGFANTWKAQSTCHDKLDWLDDPCSLNIESA   581
       * *  *** *   *         **    **  *    * * *  *   **  * *****     ** **

MUC5B  NYARHWCSRLTRSNSALSRCHSIINPKPFHSNCMFDTCNCERSEDCLCAAL--SSYVHACAAKGVQLSDW   682
MUC5AC KYAQHWCSQLTDADGPFGRCHAAVKPGTYYSNCMFDTCNCERSEDCLVRRAVLLR-ARLC-AKGVQLGGW   692
MUC2   NYAEHWCSLLKKTETPFGRCHSAVDPAEYYKRCKYDTCNCQNNEDCLCAAL--SSYARACTAKGVMLWGW   649
       *         *    ***   ***    **** **     ***    ****  ******      *  *

MUC5B  RDGVCTKYMQNCPKSQRYAYVVDACQPTCRGLSEADVTCSVSFVPVDGCTCPAGTFLNDAGACVPAQKCP   752
MUC5AC RDGVCTKPMITCPKSMTYHYHVSACQPTCRSLSEGDITCSVGFIPVDGCICPKGTFLDDTGKCVQASNCP   762
MUC2   REHVCNKDVGSCPNSQVFLYNLTTCQQTCRSLSEADSHCLEGFAPVDGCGCPDHTFLDEKGRCVPLAKCS   719
        **  *  *    * * *   * *  *       *  * ***           *   **     ** * *

MUC5B  CYAHGTVLAPGEVVHDEGAVCSCTGGKLSCLGASLQKSTGCAAPMVYLDCSNSSAGTLGAECDGSCHTLD   822
MUC5AC CYHRGSMIPNGESVHDSGAICTCTHGKLSCIGGQAP-APVCAAPMVFFDCRNATPRGTGAGCQKSCHTLD   831
MUC2   CYHRGLYLEAGDVVVRQEERCVCRDGRLHCRQIRLI-GQSCTAPKIHMDCSNLTALATSKPRALSCQTLA   788
         **   *   ** ** *     *  * *  *  *      *  **    *  **  *** *    *  *

MUC5B  VGCFSTHCVSGCVCPPGLVSDGSGGCIAEEDCPCV-NKEATYKLGETTRFDCNTCTCRNRTWECSHRLCL   891
MUC5AC MTCYSPQCVPGCVCPDGLVADGEGGCITAEDCPCVHNKA-SYRAGQTIRVGCNTCTCDSRMWRCTDDPCL   900
MUC2   AGYYHTECVSGCVCPDGLMDDGRGGCVVEKECPCVHNND-LYSSGAKIKVDCNTCTCKRGRWVCTQAVCH   857
        *****   *     *  *       * ***    * *        * ***        *    *    *

MUC5B  GTCVAYGDGHFITFDGDRYSFEGSCEYILAQDYCG-DNTTHGTFRIVTENIPCGTTGTTCSKAIKLFVES   960
MUC5AC ATCAVYGDGHYLTFDGQSYSFNGDCEYTLVQNHCGGKDSTQDSFRVVTENVPCGTTGTTCSKAIKIFLGG   970
MUC2   GTCSIYGSGHYITFDGKYYDFDGHCSYVAVQDYCG-QNSSLGSFSIITENVPCGTTGVTCSKAIKIFMGR   926
       *      *  **       *     *  ** **    *** ** ***   *      *       *  *

MUC5B  YELILQEGTFKAVARGPGGDPPYKIRYMGIFLVIETHGMAVS-WDRKTSVFIRLHQDYKARVCGLCGNFD   1029
MUC5AC FELKLSHRKVEVIGTDESQEVPYTIRQMGIYLVVDTDIGLVLLWDKKTSIFINLSPEFKGRVCGLCGNFD   1040
MUC2   TELKLEDKHRVVIQRDEGHHVAYTTREVGQYLVVESSTGIIVIWDKRTTVFIKLAPSYKGTVCGLCGNFD   996
          *           ***  ** **  * **  ***  *      ** **     * * **

MUC5B  DNAINDFATRSRSVVGDALEFGNSWKLSPSCPDALAPKDPCTANPFRKSWAEKQCSILHGPTFAACRSQV   1099
MUC5AC DIAVNDFATRSRSVVGDVLEFGNSWKLSPSCPDALAPKDPCTANPFRKSWAQKQCSILHGPTFAACHAHV   1110
MUC2   HRSNNDFTTRDHMVVSSELDFGNSWKEAPTCPDVSTNPEPCSINPHRRSWAEKQCSILKSSVFSICHSKV   1066
       * *    *  ***  **  *      ** *   ******  **  * *   *      **** ** **

MUC5B  DSTKYYEACVNDACACDSGGDCECFCTAVAAYAQACHDAGLCVSWRTPDTCPLFCDFYNPHGGCEWHYQP   1169
MUC5AC EPARYYEACVNDACACDSGGDCQCFCTAVARYAQACHEVGTCVCVRTPSICPLFCDYYNCEGQCEWHYQP   1180
MUC2   DPKPFYEACVHDSCSCDTGGDCECFCSAVASYAQECTKEGACVFWRTPDLCPIFCDYYNPPHECEWHYEP   1136
       **  *     * * *  *    *   *       * *       *   *       *  * *      *

MUC5B  CGAPCLKTCRNPSGHCLVDLPG-LERCYPKCPPSQPFNE-DQMKCVAQ--C-------GCYDKDGNYYDV   1228
MUC5AC CGVPCLRTCRNPRGDCLRDVRG-LEGCYPKCPPEAPIFDEDKMQCVAT--CPTPPLPPRCHV-HGKSYRP   1246
MUC2   CGNRSFETCRTINGIHSNISVSYLEGCYPRCPKDRPIYEEDLKKCVTADKC-------GCYV-EDTHYPP   1199
          ***    **   ** *  *   *   *      * **  **  *           * **  *    *

MUC5B  GARVPTAENCQSCNCTPSGIQ-CAHSLEACTCTYEDRTYSYQDVIYNTTDGLGACLIAICGSNGTIIRKA   1297
MUC5AC GAVVPSDKNCQSCLCTERGVE-CTYKAEACVCTYNGQRFHPGDVIYHTTDGTGGCISARCGANGTIERRV   1315
MUC2   GASVPTEETCKSCVCTNSSQVVC---------------RPEEGKILNGTQDGAFCYWEICGPNGTVEKHF   1254
            * ** *      **        *** ***        ** *** ** *    **      ***

MUC5B  VACPGTPATTPFTFTTAVPHSTTS                                                 1321
MUC5AC YPCSPTTPVPPTTTSFSTPPLVVS                                                 1339
MUC2   NICSITTRPSTLTTFTTITLPTTP                                                 1278
          *  *   *  * *  *  ***
```

FIG. 2. Comparison of the deduced N-terminal sequence of MUC5B with that of MUC5AC and MUC2. Cysteine residues are indicated by dark gray boxes and all other residues identical in all three sequences are indicated by light gray boxes. Residues identical in two of the three sequences are indicated with asterisks. The position of the putative signal peptide cleavage site is marked with an arrow. The sequences of MUC5AC and MUC2 were derived from (16, 17) and (10, 11), respectively.

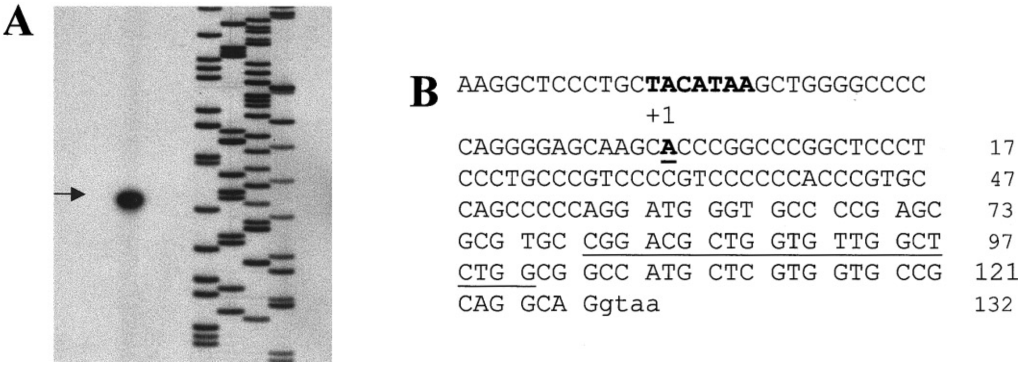

A
B
AAGGCTCCCTGCTACATAAGCTGGGGCCCC
+1
CAGGGGAGCAAGCACCCGGCCCGGCTCCCT 17
CCCTGCCCGTCCCCGTCCCCCCACCCGTGC 47
CAGCCCCCAGG ATG GGT GCC CCG AGC 73
GCG TGC CGG ACG CTG GTG TTG GCT 97
CTG GCG GCC ATG CTC GTG GTG CCG 121
CAG GCA Ggtaa 132

FIG. 3. Identification of the transcription start site in the MUC5B gene by primer extension. (A) The primer extension product obtained using RNA (50 μg) from human gallbladder epithelial cells and the primer underlined in (B) was electrophoresed on a 6% acrylamide 7 M urea sequencing gel next to an M13 sequencing ladder obtained using the −40 universal primer. The lanes were loaded in the order GATC. The position of the major 101-base product is indicated with an arrow. (B) Sequence surrounding the transcription start site in the MUC5B gene. The putative TATA box element is indicated in bold-faced type and the "A" (+1) identified as the transcription start site by primer extension is underlined. The primer (SP7) used in the primer extension reaction is also underlined. The exon/intron boundary near the end of the signal peptide is indicated by capital/lower case letters.

the three mucins reveals that the D3 domains display a higher degree of identity than do the D1 and D2 domains. The presence of CGLCG sequences in all D3 domains, together with the high degree of sequence identity, suggests that these domains may have a role in mucin polymerization similar to that in pre pro-von Willebrand factor.

It is noteworthy that the N-terminal MUC5B D domains contain structural motifs not previously recognized in mammalian gel-forming mucins. First, an IDENTIFY pattern search revealed consensus motifs for C-type lectin domain proteins and selectin complement binding repeats in the D3 domain. C-type lectin domains are found on a wide array of extracellular proteins and have been shown to bind calcium, protein ligands and carbohydrates (26, 27). The selectin complement binding repeats have been shown to mediate the interactions between P-selectin and leukocytes (28). These presence of these motifs suggested that the N-terminal D3 domain of MUC5B may also have an important role in ligand binding or adhesion. A search of GenBank revealed that all three N-terminal MUC5B D domains contained sequences which were similar to cysteine-rich EGF repeats in the extracellular proteins laminin and tenascin and calcium binding EGF repeats in human fibrillin and in the Drosophila neurogenic notch protein. These repeats are present in many extracellular proteins including cell surface receptors and extracellular matrix components (29). The function and ligand binding properties of EGF-like domains in many of these molecules is incompletely understood, although in several instances, these sequences have been shown to mediate specific receptor-type protein-protein interactions (30, 31). In addition, each of the D domains contained motifs which were similar to a block of 4 cysteine-rich LDL receptor Class A repeats present in the LDL receptor family proteins megalin and the a_2 macroglobulin receptor. The LDL receptor repeats in megalin have been shown to bind a wide variety of ligands in vitro including apoliprotein J, plasminogen, lipoprotein lipase, lactoferrin and calcium as well as polycationic drugs (32).

We have previously shown that bovine gallbladder mucin contains cysteine-rich sequences similar to the scavenger receptor cysteine-rich repeats present in many ligand binding proteins (33) and have demonstrated that domains containing these repeats bind biliary lipids in vitro (34). The presence of multiple binding domains in the N-terminal region of MUC5B suggests a mechanism by which this mucin could perform different functions in different biological environments. In the gallbladder, such domains could be involved in the interactions of mucin with other biliary proteins or lipids, possibly playing a role in the pathogenesis of gallstones. In the oral cavity, such domains could be involved in interactions of mucins with other salivary proteins and in the binding and clearance of microbes. In the trachea and reproductive tract, MUC5B binding domains could also play a significant role in physiological processes such as non-immune host defense. Finally, the present work indicates that in MUC5B, as well as in other gel forming mucins, the D domains may exhibit multifunctional properties necessary for both mucin assembly and a variety of tissue-specific functions.

ACKNOWLEDGMENTS

This work was supported in part by NIH Grants DK44619, DE11691, and DK 02327.

REFERENCES

1. Gendler, S. J., and Spicer, A. P. (1995) *Ann. Rev. Physiol.* **57,** 60–634.
2. Van Klinken, B. J. W., Dekker, J., Buller, H. A., and Einerhand, A. W. (1995) *Am. J. Physiol.* **269,** G613–G627.
3. Dufosse, J., Porchet, N., Audie, J. P., Guyonnet Duperot, V., Laine, A., Van-Seuningen, I., Marrakchi, S., Degand, P., and Aubert, J. P. (1993) *J. Biol. Chem.* **293,** 329–337.
4. Keates, A. C, Nunes, D. P., Afdhal, N. H., Troxler, R. F., and Offner, G. D. (1997) *Biochem. J.* **324,** 295–303.
5. Troxler, R. F., Iontcheva, I. I., Oppenheim, F. G., Nunes, D. P., and Offner, G. D. (1997) *Glycobiology* **7,** 965–973.
6. Desseyn, J. L., Aubert, J. P., Van Seuningen, I., and Porchet, N. (1997) *J. Biol. Chem.* **272,** 16873–16883.
7. Neilsen, P. A., Bennett, E. P., Wanall, H. H., Therkildsen, M. H., Hannibal, J., and Clausen, H. (1997) *Glycobiology* **7,** 413–419.
8. Gipson, I. K., Ho, S. B., Spurr-Michaud, S. J., Tisdale, A. S., Zhan, Q., Torlakovic, E., Pudney, J., Anderson, D. J., Toribara, N. H., and Hill, J. A. (1997) *Biol. Reprod.* **56,** 999–1011.
9. Desseyn, J. L., Buisine, M. P., Porchet, N., Aubert, J. P., Degand, P., and Laine, A. (1998). *J. Mol. Evol.* **46,** 102–106.
10. Gum, J. R., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E., and Kim, Y. S. (1992) *J. Biol. Chem.* **267,** 21375–21383.
11. Gum, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) *J. Biol. Chem.* **269,** 2440–2446.
12. Sadler, J. E. (1991) *J. Biol. Chem.* **266,** 22777–22780.
13. Meerzaman, D., Charles, P., Daskal, E., Polymeropopoulos, M. H., Martin, B. M., and Rose, M. C. (1994) *J. Biol. Chem.* **269,** 12932–12939.
14. Lesuffleur, R., 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.
15. Toribara, N. W., Ho, S. B., Gum, E., Gum, J. R., Lau, P., and Kim, Y. S. (1997) *J. Biol. Chem.* **272,** 16398–16403.
16. Klomp, L. W. J., Van Rens, L., and Strous, G. J. (1995) *Biochem. J.* **308,** 831–838.
17. Li, D., Gallup., M., Fan, N., Szymkowski, D. E., and Basbaum, C. B. (1998) *J. Biol. Chem.* **273,** 6812–6820.
18. van de Bovenkamp, J. H. B., Hau, C. M., Strous, G. J. A. M., Buller, H. A., Dekker, J., and Einerhand, A. W. C. (1998) *Biochem. Biophys. Res. Commun.* **245,** 853–859.
19. Chomzynski, P., and Sacchi, N. (1987) *Anal. Biochem.* **162,** 156–165.
20. Purdum. P. P., Ulissi, A., Hylemon, P. B., Shiffman, M. L., and Moore, E. W. (1993) *Lab. Invest.* **68,** 345–353.
21. Desseyn, J. L., Guyonnet-Duperat, V., Porchet, N., Aubert, J. P., and Laine, A. (1997) *J. Biol. Chem.* **272,** 3168–3178.
22. Kozak, M. (1991) *J. Biol. Chem.* **266,** 19867–19870.
23. Dong, Z., Thoma, R. S., Crimmins, D. L., McCourt, D. W., Tuley, E. A., and Sadler, J. E. (1994) *J. Biol. Chem.* **269,** 6753–6758.
24. Mayadas, T. N., and Wagner, D. D. (1992) *Proc. Nat. Acad. Sci. (US)* **89,** 3531–3535.
25. Perez-Vilar, J., Eckhardt, A. E., DeLuca, A., and Hill, R. L. (1998) *J. Biol. Chem.* **273,** 14442–14449.
26. Aspberg, A., Miura, R., Bourdoulous, S., Shimonaka, M., Heingard, D., Schachner, M., Ruoslahti, E., and Yamaguchi, Y. (1997) *Proc. Nat. Acad. Sci. (US)* **94,** 10116–10121.
27. Hosoi, T., Imai, Y., and Irimura, T. (1998) *Glycobiology* **8,** 791–798.
28. Ruchaud-Sparagano, M.H., Malaud, E., Gayet, O., Chignier, E., Buckland, R., and McGregor, J. L. (1998) *Biochem. J.* **332,** 309–314.
29. Bork, P. (1991) *FEBS Lett.* **286,** 47–54.
30. Mayer, U., Nischt, R., Poschl, E., Mann, K., Fukuda, K., Gerl, M., Yamada, Y., and Timpl, R. (1993) *EMBO J.* **12,** 1879–1885.
31. Rand, M. D., Lindblom, A., Carlson, J., Villoutreix, B. O., and Stenflo, J. (1997) *Protein Sci.* **6,** 2059–2071.
32. Orlando, R. A., Exner, M., Czekay, R. P., Yamazaki, H., Saito, A., Ullrich, R., Kerjaschki, D., and Farquhar, M. G. (1997) *Proc. Nat. Acad. Sci. (US)* **94,** 2368–2373.
33. Nunes, D. P., Keates, A. C., Afdhal, N. H., and Offner, G. D. (1995) *Biochem. J.* **310,** 41–48.
34. Nunes, D. P., Afdhal, N. H., Niu, N., Keates, A. C., and Offner, G. D. (1994) *Gastroenterology* **106,** 951A.