BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **238,** 143–148 (1997)
ARTICLE NO. RC977258

Molecular Cloning of Human MUC3 cDNA Reveals a Novel 59 Amino Acid Tandem Repeat Region[1]

B. Jan-Willem Van Klinken,* Tanja C. Van Dijken,* Esmee Oussoren,* Hans A. Büller,† Jan Dekker,* and Alexandra W. C. Einerhand*,[2]

**Pediatric Gastroenterology and Nutrition, Emma Children's Hospital AMC, University of Amsterdam, Amsterdam, The Netherlands; and †Sophia Children's Hospital, Rotterdam, The Netherlands*

Received August 4, 1997

The human MUC3 gene is highly expressed in small intestine and gallbladder. Thus far only 646 basepairs of its cDNA encoding 17 amino acid repeats have been cloned. In order to further clone the human MUC3 cDNA, a human small intestinal cDNA library was constructed and screened with a cDNA probe encompassing the 17 amino acid tandem repeat region of human MUC3. In two subsequent screenings of the library resulting positive clones were used as probes. In total, 27 partial MUC3 cDNA clones were isolated and sequenced that define a semi-unique region and a novel 177 nucleotide tandem repeat region, located upstream of the region encoding the 17 amino acid tandem repeats. The 177 nucleotide repeat region is at least 5 kb in length and encodes 59 amino acid repetitive peptides with a consensus sequence of VSTTPVASSEASTLSTTPVDTSTPVTTSTQASSSPTTAEGTSMPTSTPSEGSTPLTSMP, that is notably different from the 17 amino acid repeat of MUC3 or any other known mucin repeat. © 1997 Academic Press

Epithelial mucins are very large glycoproteins that are encoded by a family of at least nine human genes, designated MUC1-4,5AC, 5B, and MUC6-8 (see for review 1). Reported sequences of the corresponding cDNAs are rarely full-length, because of the repetitive nature and the extremely large size of the mucin cDNAs. Thus far, the full-length cDNAs sequences of the relatively small human mucins, MUC1 and MUC7, have been analyzed (2,3). In addition, the large full-length MUC2 cDNA, which is over 15 kb in length, has been completely cloned and sequenced (4-6). Analysis of the central part of full-length cDNA sequences of MUC1,2 and 7 revealed many repetitive serine, threonine and proline-rich sequences that differed in sequence and in length. The partial cDNA analysis of MUC3,4, 5AC, 5B, 6 and 8 also revealed that each of these mucins contained distinct repetitive regions (7-12).

We are interested in studying the expression and function of intestinal mucin genes. MUC2 and MUC3 genes are highly expressed in the human small intestine. The MUC2 gene in addition to its full-length cDNA has been cloned (13). Moreover, its biosynthesis and secretion in the intestine has been well characterized (14). However, only 646 bp of the MUC3 cDNA encompassing a 51 nucleotide repeat region has been cloned and sequenced (7). Thus far the available MUC3 antibodies are raised against the repetitive polypeptide backbone of MUC3. These antibodies do not recognize the mature MUC3 glycoprotein (15,16), most likely because the repeat region is heavily glycosylated. Therefore, the biosynthesis, intracellular processing and localization of mature MUC3 could not be studied. In order to obtain more cDNA sequences of MUC3 and thus be able to raise antibodies against the unique non-repetitive sequences of MUC3, we sought to clone and characterize the MUC3 regions flanking the 17 amino acid tandem repeat region. We constructed and subsequently screened a human small intestinal cDNA library with a cDNA probe encompassing the 17 amino acid tandem repeat region of human MUC3 and report here the identification of a novel large repetitive region located upstream of the 17 amino acid repeat region, that is distinct from any other known mucin repeat.

MATERIALS AND METHODS

Construction and screening of a human small intestinal cDNA library. Human jejunal tissue was resected from a patient undergo-

[1] The nucleotide sequences presented in this paper have been submitted to Genbank with accession numbers: AF016692, AF016693, and AF016694.

[2] Corresponding author: Alexandra W. C. Einerhand, Academic Medical Center, Room G8-205, Pediatric Gastroenterology and Nutrition, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. Fax: 31-20-6917735. E-mail: a.w.einerhand@amc.uva.nl.

Abbreviations: GITC, guanidinium isothiocyanate; Endo-H, endo-β-N-acetylglucoaminidase H.

0006-291X/97 $25.00

ing a Whipple operation. The tissue was healthy as judged by macroscopic criteria. Total RNA was isolated from human jejunal scrapings after dissolution in guanidium isothiocyanate (GITC) and centrifugation on a CsCl cushion as described earlier (15). Poly-A^+-RNA was subsequently isolated using the poly-A^+-tract system following the manufacturers protocol (Promega). Five μg poly-A^+-RNA was used to synthesize cDNA with the Superscript system (Gibco-BRL) following the manufacturers description using a mixture of oligo-dT and random hexamer primers. EcoRI-adaptors were ligated to the cDNA ends, whereafter the cDNA was size-fractionated using a Sephadex S200 column. Fractions were pooled containing cDNAs larger than 1 kb, and ligated into the EcoRI site of the lambda ZAPII vector (Stratagene) and packaged using the Gigapack Gold kit (Stratagene). The resulting library contained 7.5×10^7 independent plaque forming units (pfu). Part of the non-amplified library (1×10^5 pfu) was plated onto *E. coli* XL1-Blue and screened according to the manufacturers protocol (Stratagene) initially with a ^{32}P-labeled EcoRI cDNA-fragment derived from the published clone SIB139 containing the 51 bp MUC3 repeats (7). In a subsequent screening a ^{32}P-labeled EcoRI-NcoI fragment of one of the resulting positive cDNA clones, pMUC3T30, was used as probe. In the final screening of the library, a ^{32}P-labeled AccI fragment of one of the resulting clones, pMUC3T26, of the second screening was used as probe. Filters (Schleicher and Schuell) were hybridized in duplicate at 65°C in 0.5 M NaH_2PO_4 pH 7.2, 7% SDS and 1 mM EDTA according to the method of Church and Gilbert (17). Positive plaques were rescreened and the corresponding plasmids were excised with the help of Exassist helper phage according the manufacturers protocol (Stratagene). Plasmids were isolated using the Wizard Plus kit (Promega) and subsequently sequenced.

Sequence analysis of cDNA clones. MUC3 cDNA containing plasmids, pMUC3T3,5,9,10,11,21,26,61, were double-strandedly sequenced using Taq dye-nucleotide cycle sequencing kit with fluorescently labeled nucleotides (Applied Biosystems) and T3 and T7 primers or MUC3 specific primers 5′-CTGTCCTCATCAGCCC-3′ and 5′-GTCACATATGTGAGGGG-3′ in a Perkin Elmer 9600 thermocycler. The MUC3 cDNA inserts of plasmids pMUC3T4,6,12,15,19,51,55, 56,63,65,68 were partially and only single-strandedly sequenced from both ends of the inserts using T3 and T7 primers following the same sequence method as mentioned above. Sequence reactions were analyzed on an Applied Biosystems model 377 sequencer. Sequences were analyzed using Macintosh Sequence Navigator and Autoassembler software.

Dot-blot analysis. Total RNA was isolated from mucosal scrapings of resected, healthy segments of gallbladder, stomach, and small and large intestine. These scrapings were homogenized in GITC. In addition, a GITC lysate of a primary human trachael cell culture was kindly provided by dr P. Nettesheim. RNA was isolated from the GITC-lysates using ultracentrifugation on a CsCl cushion (15). Integrity of all RNAs was assessed by analysis of the 28S and 18S ribosomal RNAs after electrophoresis on a 0.8% agarose gel and staining by ethidium bromide. Dot-blot analysis was essentially carried out as described earlier (15). In short, 1 μg RNA from each tissue was dot blotted in duplicate onto N-Hybond membranes using a vacuum-operated dot-blot apparatus (Biorad). Membranes were baked for 2 h at 80°C and subsequently hybridized according to the method of Church and Gilbert (17) with either a ^{32}P-labeled EcoRI cDNA-fragment derived from the published clone SIB139 containing the 51 bp MUC3 repeats (7) or with the ^{32}P-labeled NcoI-EcoRI fragment derived from plasmid pMUC3T30. Radioactive signals on the membranes were quantified with a PhosphorImager using ImageQuant software (Molecular Dynamics, Sunnyvale, CA). The membranes were then washed to remove the mucin probes, checked for residual activity, and hybridized to ^{32}P-labeled β-actin as a measure of the amount of RNA blotted as described earlier (15). The MUC3 radioactive signals were expressed as ratio relative to the β-actin signal to standardize for the amount of RNA blotted. The use of human tissue was approved by the medical ethical committee of our institution.

RESULTS

Isolation of MUC3 cDNA Clones Containing 51 bp Repeat and Semi-unique Sequences

Human MUC2 and MUC3 are important intestinal mucins that are both very highly expressed in the small intestine. In contrast to MUC2, however, not much is known about the MUC3 cDNA sequence, or the biosynthesis, intracellular processing and the subsequent localization of the mature MUC3 glycoprotein. The fact the MUC3 research lags behind that of MUC2 can at least in part be explained by the lack of antibodies that recognize the fully glycosylated MUC3 protein. Thus far only 646 bp of a 51 bp tandem nucleotide repeat of the MUC3 cDNA has been cloned and sequenced (7). The region only encodes a 17 amino acid repeat with the consensus sequence HSTPSFTSSITTTETTS, containing many potential *O*-glycosylation sites.

In order to further study MUC3 and be able to raise antibodies recognizing the mature MUC3, we have sought to clone the unique non-repetitive sequences of human MUC3 flanking the 17 amino acid repetitive region. Therefore, we constructed a human small intestinal lambda ZAPII cDNA library containing 7.5×10^7 independent recombinant plaque forming units. Approximately 100.000 recombinant plaques were screened with the MUC3 51 bp repeat probe. Six independent plaques were positive and sequence analysis revealed that five of them contained inserts that completely consisted of repeats similar, but not identical (85%-99%) to the published 51 bp MUC3 repeat sequence (7). The inserts were on average 800 bp in length. None of these five clones were either identical or overlapping in sequence (data not shown).

In contrast to the five clones containing only 51 bp repeat sequences, one of the six positive plaques, named pMUC3T30, contained a 1.2 kb insert consisting not only of the 51 bp repeats, but also of a non-repetitive region encoding many serines and threonines located upstream of the 51 bp repeat (Fig. 2A).

Dot-blot Analysis with the 5′-region of pMUC3T30 as Probe

To ensure that the newly identified sequence of clone pMUC3T30 is an intrinsic part of the MUC3 cDNA, a RNA dot-blot analysis was carried out using an EcoRI-NcoI fragment containing the 5′-end of the insert pMUC3T30 as probe (Fig. 2A, probe A). A previous RNA dot-blot analysis using the 51 bp repeat sequences as probe, showed that MUC3 is highly expressed in small intestine and gallbladder, whereas expression was hardly detectable in trachea, stomach and colon (15). The RNA dot-blot analysis shown in figure 1 demonstrates that the signals derived from the 51 bp repeat and the newly identified sequence are comparable. In particular, the signals in small intestine compared to

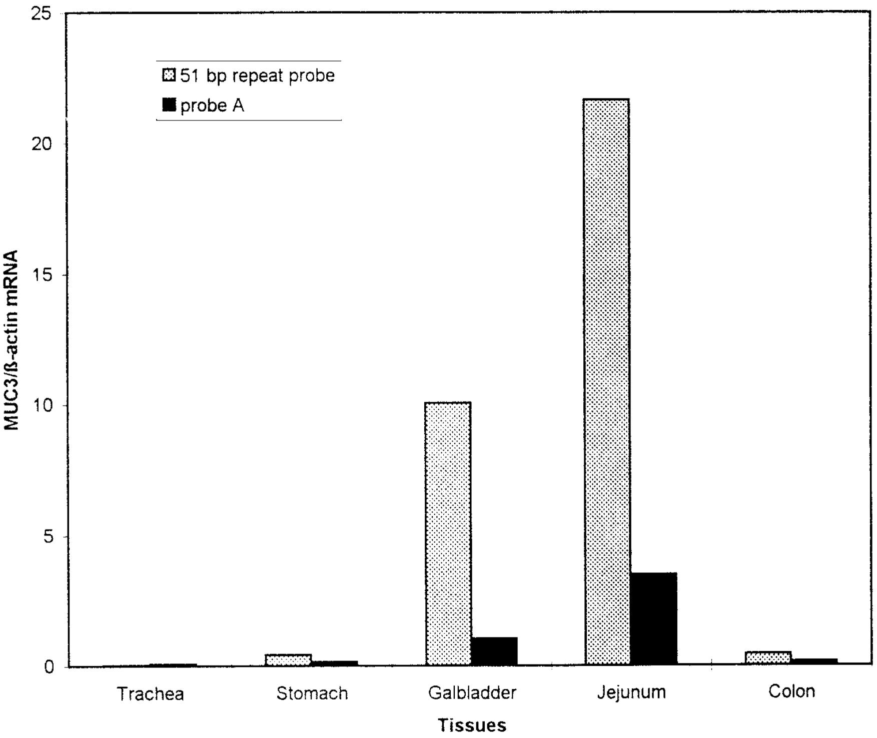


FIG. 1. Tissue RNA dot-blot analysis. Total RNA derived from indicated epithelia was dot-blotted and hybridized to either a radioactively labeled SIB139 cDNA probe (7) containing the 51 bp tandem repeat of MUC3 or to a radioactively labeled EcoRI-NcoI cDNA fragment of pMUC3T30 corresponding to probe A shown in figure 2A. To ensure that equal amounts of RNA were dot-blotted the blots were stripped of the MUC3 probes and reprobed with a radioactively labeled β-actin probe. MUC3 radioactive signals were quantified using a PhosphorImager and expressed relative to the signal elicited by β-actin.

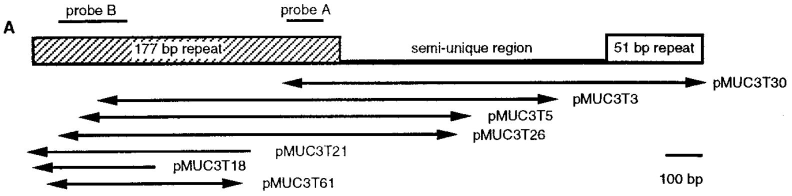
gallbladder for both probes was more than 2-fold higher, whereas signals in trachea, stomach and colon were hardly detectable, indicating that the newly identified sequence most likely hybridized with the MUC3 mRNA.

Isolation of MUC3 cDNA Clones Containing a Novel 177 bp Repeat

To clone additional MUC3 sequences the EcoRI-NcoI fragment (probe A, Fig. 2A) of clone pMUC3T30 was subsequently used to screen approximately 100.000 recombinant plaques of the non-amplified human intestinal cDNA library. After rescreening 5 positive plaques were isolated and the sequences of the corresponding inserts were analyzed. Three clones, pMUC3T3,5 and 26 contained sequences overlapping with the sequences of the probe (Fig. 2A). Surprisingly, analysis of the 5′-ends of these sequences revealed the presence of a novel 177 bp degenerate tandem repeat (Fig. 2B). The inserts of the 2 other clones, pMUC3T4 and pMUC3T25, were respectively 2.7 kb and 1.5 kb in length and sequenced from both ends (data not shown). These sequences were very similar but not identical to the 177 bp repeat sequences of clones PMUC3T3,5 and 26, indicating that the 177 bp repeat region is at least 3 kb in length. However, still no non-repetitive MUC3 sequence outside of the repeat region was cloned. Therefore, the most 5′-end, a radioactively labeled AccI-fragment, of pMUC3T26 containing the novel 177 bp repeat was used as probe (Fig. 2B, probe B) to screen 100.000 recombinant plaques of the human small intestinal cDNA library. Three clones, pMUC3T18, 21 and 61 overlapped with the sequence of the probe containing the 177 bp repeat (Fig. 2A). The sequences of all overlapping clones shown in figure 2A were combined and this resulted in the identification of about 1.8 kb sequence upstream of the 51 bp repeat (Fig 2B). In addition, 13 clones, pMUC3T6,9,10,11,12,15,19, 51,55,56,63,65,68 were obtained that did not overlap with the sequence shown in figure 2B. However, each of 13 corresponding inserts contained 177 bp repetitive sequences similar, but not identical to the 177 bp repeat used as probe. As representative examples two of these double-strandedly sequenced clones, pMUC3T9 and pMUC3T10 are shown in figures 2C and 2D.

Analysis of MUC3 Amino Acid Sequences

Translation of the sequence shown in figure 2B revealed 5 degenerate 59 amino acid repeats (Fig. 3C, position 1-295) at the 5′-end and 5 repeats of 17 amino



B

```
       177 bp repeat
   1 CTCTCAGTAC CACGCCGGTG GCCAGTTCTG AGGCTAGCAC CCTTTCAACA AGTCCTGTTG ACACCAGCAC ACCTGTGACC
  81 AATTCTTCTC CAACCAATTC ATCTCCTACA ACTGCTGAAG TTACCAGCAT GCCAACATCA ACTGCTGGTG AAGGAAGCAC
 161 TCCATTAACA AATATGCCTG TCAGCACCAC ACCGGTGGCC AGTTCTGAGG CTAGCACCCT TTCAACAACT CCTGTTGACT
 241 CCAACACTTT TGTTACCAGT TCTAGTCAAG CCAGTTCATC TCCAGCAACT CTTCAGGTCA CCACTATGCG TATGTCTACT
 321 CCAAGTGAAG GAAGCTCTTC ATTAACAACT ATGCTCCTCA GCAGCACATA TGTGACCAGT TCTGAGGCTA GCACACCTTC
 401 CACTCCTTCT GTTGACAGAA GCACACCTGT GACCACTTCT ACTCAGAGCA ATTCTACTCC TACACCTCCT GAAGTTATCA
 481 CCCTGCCAAT GTCAACTCCT AGTGAAGTAA GCACTCCATT AACCATTATG CCTGTCAGCA CCACATCGGT GACCATTTCT
 561 GAGGCTGGCA CAGCTTCAAC ACTTCCTGTT GACACCAGCA CACCTGTGAT CACTTCTACC CAAGTCAGTT CATCTCCTGT
 641 GACTCCTGAA GGTACCACCA TGCCAATCTG GACGCTTAGT GAAGGAAGCA TTCCCATTAC AATTATGCAT GTCAGCACCA
 721 CACGTGTGAC CAGTTCTGAG GGTAGCACCC TTTCAACACC TTCTGTTGTC ACCAGCACAC CTGTGACCAC TTCTACTGAA
 801 GCCATTTCAT CTTCTGCAAC TCTTGACAGC ACCACCATGT CTGTGTCAAT GCCCATGGAA ATAAGCACCC TTGGGACCAC
                semi-unique region
 881 TATTCTT[GTC AGTACCACAC CTGTTACGAG GTTTCCTGAG AGTAGCACCC CTTCCATACC ATCTGTTTAC ACCAGCATGT
 961 CTATGACCAC TGCCTCTGAA GGCAGTTCAT CTCCTACAAC TCTTGAAGGC ACCACCACCA TGCCTATGTC AACTACGAGT
1041 GAAAGAAGCA CTTTATTGAC AACTGTCCTC ATCAGCCCTA TATCTGTGAT GAGTCCTTCT GAGGCCAGCA CACTTTCAAC
1121 ACCTCCTGGT GATACCAGCA CACCTTTGCT CACCTCTACC AAAGCCGGTT CATTCTCCAT ACCTGCTGAA GTCACTACCA
1201 TACGTATTTC AATTACCAGT GAAAGAAGCA CTCCATTAAC AACTCTCCTT GTCAGCACCA CACTTCCAAC TAGCTTTCCT
1281 GGGGCCAGCA TAGCTTCGAC ACCTCCTCTT GACACAAGCA CAACTTTTAC CCCTTCTACT GACACTGCCT CAACTCCCAC
1361 AATTCCTGTA GCCACCACCA TATCTGTATC AGTGATCACA GAAGGAAGCA CACCTGGGAC AACCATTTTT ATTCCCAGCA
1441 CTCCTGTCAC CAGTTCTACT GCTGATGTCT TTCCTGCAAC AACTGGTGCT GTATCTACCC CTGTGATAAC TTCCACTGAA
1521 CTAAACACAC CATCAACCTC CAGTAGTAGT ACCACCACAT CTTTTTCAAC TACTAAGGAA TTTACAACAC CCGCAATGAC
1601 TACTGCAGCT CCCCTCACAT ATGTGACCAT GTCTACTGCC CCCAGCACAC CCAGAACAAC CAGCAGAGGC TGCACTACTT
          semi-unique region      51 bp  repeat
1681 CTGCATCAAC GCTTTCTGCA]ACCAGTACAC CTCACACCTC TACTTCGATT ACCACCACCG AGACCCCCTC AAGCAGTACT
1761 CCCAGCTTCA CTTCTTCGAT CACCACCACC GAGACCACAT CCCACAGTAC TCCCAGCTTC ACTTCTTCAA TCAGCACCAC
1841 TGAGACCACA TCCCACAGTA CTCCCAGCTT CACTTCTTTG ATCACCATCA CCGAGACCAC CTCACACAGT ACTCCCAGCT
1921 ACACTACCTC AATCACCACC ACCGAGC
```

C

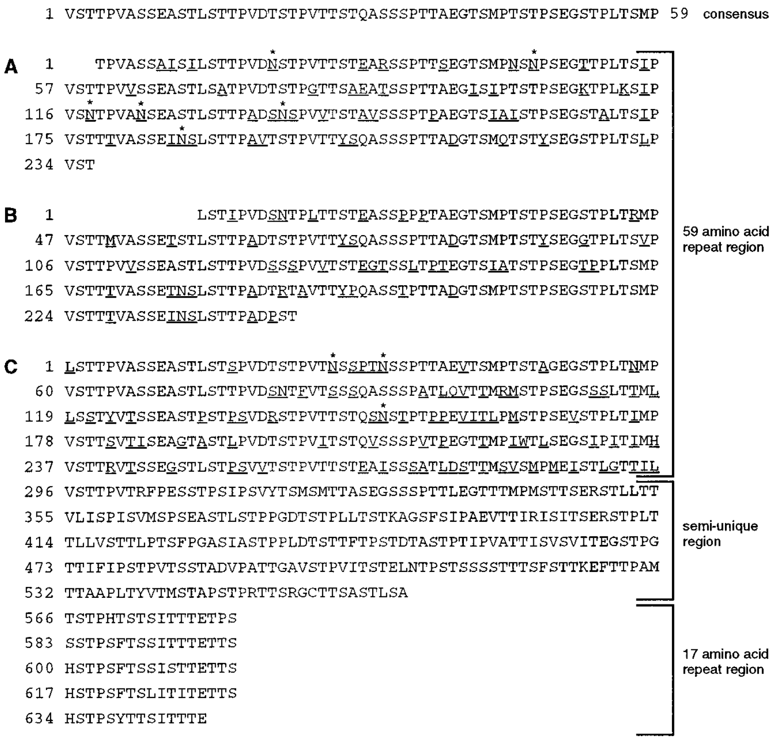
```
   1 CACACCGGTG GCCAGTTCTG CAATCAGCAT CCTTTCAACA ACTCCTGTTG ACAACAGCAC ACCTGTGACC ACTTCTACTG
  81 AAGCCCGTTC ATCTCCTACA ACTTCTGAAG GTACCAGCAT GCCAAACTCA AATCCTAGTG AAGGAACCAC TCCGTTAACA
 161 AGTATACCTG TCAGCACCAC GCCGGTAGTC AGTTCTGAGG CTAGCACCCT TTCAGCAACT CCTGTTGACA CCAGCACCCC
 241 TGGGACCACT TCTGCTGAAG CCACTTCATC TCCTACAACT GCTGAAGGTA TCAGCATACC AACCTCAACT CCTAGTGAAG
 321 GAAAGACTCC ATTAAAAAGT ATACCTGTCA GCAACACGCC GGTGGCCAAT TCTGAGGCTA GCACCCTTTC AACAACTCCT
 401 GCTGACTCTA ACAGTCCTGT GGTCACTTCT ACAGCAGTCA GTTCATCTCC TACACCTGCT GAAGGTACCA GCATAGCAAT
 481 CTCAACGCCT AGTGAAGGAA GCACTGCATT AACAAGTATA CCTGTCAGCA CCACAACAGT GGCCAGTTCT GAAATCAACA
 561 GCCTTTCAAC AACTCCTGCT GTCACCAGCA CACCTGTGAC CACTTATTCT CAAGCCAGTT CATCTCCTAC AACTGCTGAC
 641 GGTACCAGCA TGCAAACCTC AACTTATAGT GAAGGAAGCA CTCCACTAAC AAGTTTGCCT GTCAGCACCC
```

D

```
   1 CCTTTCAACA ATTCCTGTTG ACTCCAACAC TCCTTTGACC ACTTCTACTG AAGCCAGTTC ACCTCCTCCC ACTGCTGAAG
  81 GTACCAGCAT GCCAACCTCA ACTCCTAGTG AAGGAAGCAC TCCATTAACA CGTATGCCTG TCAGCACCAC AATGGTGGCC
 161 AGTTCTGAAA CGAGCACACT TTCAACAACT CCTGCTGACA CCAGCACACC TGTGACCACT TATTCTCAAG CCAGTTCATC
 241 TCCTACAACT GCTGACGGTA CCAGCATGCC AACCTCAACT TATAGTGAAG GAGGCACTCC ACTAACAAGT GTGCCTGTCA
 321 GCACCACGCC GGTGGTCAGT TCTGAGGCTA GCACCCTTTC AACAACTCCT GTTGACTCTA GCAGTCCTGT GGTCACTTCT
 401 ACTGAAGGCA CTTCATCTCT TACACCTACT GAAGGTACCA GCATAGCAAC CTCAACGCCT AGTGAAGGAA CACCTCCATT
 481 AACAAGTATG CCTGTCAGCA CCACAACAGT GGCCAGTTCT GAAACCAACA GTCTTTCAAC AACTCCTGCT GACACCAGGA
 561 CAGCTGTGAC CACTTATCCT CAAGCCAGTT CAACTCCTAC AACTGCTGAC GGTACCAGCA TGCCAACCTC AACTCCTAGT
 641 GAAGGAAGCA CTCCATTAAC AAGTATGCCT GTCAGCACCA CAACAGTGGC CAGTTCTGAA ATCAACAGTC TTTCAACAAC
 721 TCCTGCTGAC CCCAGCAC
```

FIG. 2. Analysis of the isolated human MUC3 cDNA clones. (A) Schematic overview of the isolated and analyzed cDNA clones mapped upstream of the 17 amino acid tandem repeat region (open box). The 177 bp repeat (striped box) and semi-unique regions are indicated. cDNA clones single or double-strandedly sequenced are marked by arrows at one or both ends, respectively. The EcoRI-NcoI-fragment of pMUC3T30 and the AccI-fragment of pMUC3T26, which were subsequently used as probes to screen the cDNA library, are illustrated at the top of the figure as probes A and B, respectively. (B) Combined cDNA sequence derived from the clones schematically illustrated in panel A (Genbank accession number: AF016692). Arrows indicate the start of either the 177 bp repeats in the 5′-region or the 51 bp repeat in the 3′-region of the sequence. The semi-unique region is located between brackets. (C) Complete cDNA sequence of pMUC3T9 (Genbank accession number: AF016693). Arrows indicate the start of the 177 bp repeats. (D) Complete cDNA sequence of pMUC3T10 (Genbank accession number: AF016694). Arrows indicate the start of the 177 bp repeats.

acids at the 3′-end (position 566-647). In between these two repeat regions a non-repetitive semi-unique region of 270 amino acids is located (position 296-565) that is, analogous to both repeat regions, rich in serine, threonine and proline residues. At the border between the semi-unique sequence and the 17 amino acid repeat region one cysteine residue is located at position 556. Translation of the clones pMUC3T9 and pMUC3T10 revealed an additional 8 repeats of 59 amino acids resulting in the following consensus sequence VSTTPV-



```
    1 VSTTPVASSEASTLSTTPVDTSTPVTTSTQASSSPTTAEGTSMPTSTPSEGSTPLTSMP 59  consensus

A   1    TPVASSAISILSTTPVDNSTPVTTSTEARSSPTTSEGTSMPNSNPSEGTTPLTSIP
   57 VSTTPVVSSEASTLSATPVDTSTPGTTSAEATSSPTTAEGISIPTSTPSEGKTPLKSIP
  116 VSNTPVANSEASTLSTTPADSNSPVVTSTAVSSSPTPAEGTSIAISTPSEGSTALTSIP
  175 VSTTTVASSEINSLSTTPAVTSTPVTTYSQASSSPTTADGTSMQTSTYSEGSTPLTSLP
  234 VST

B   1                LSTIPVDSNTPLTTSTEASSPPPTAEGTSMPTSTPSEGSTPLTRMP
   47 VSTTMVASSETSTLSTTPADTSTPVTTYSQASSSPTTADGTSMPTSTYSEGGTPLTSVP
  106 VSTTPVVSSEASTLSTTPVDSSSPVVTSTEGTSSLTPTEGTSIATSTPSEGTPPLTSMP
  165 VSTTTVASSETNSLSTTPADTRTAVTTYPQASSTPTTADGTSMPTSTPSEGSTPLTSMP
  224 VSTTTVASSEINSLSTTPADPST

C   1 LSTTPVASSEASTLSTSPVDTSTPVTNSSPTNSSPTTAEVTSMPTSTAGEGSTPLTNMP
   60 VSTTPVASSEASTLSTTPVDSNTFVTSSSQASSSPATLQVTTMRMSTPSEGSSSLTTML
  119 LSSTYVTSSEASTPSTPSVDRSTPVTTSTQSNSTPTPPEVITLPMSTPSEVSTPLTIMP
  178 VSTTSVTISEAGTASTLPVDTSTPVITSTQVSSSPVTPEGTTMPIWTLSEGSIPITIMH
  237 VSTTRVTSSEGSTLSTPSVVTSTPVTTSTEAISSSATLDSTTMSVSMPMEISTLGTTIL
  296 VSTTPVTRFPESSTPSIPSVYTSMSMTTASEGSSSPTTLEGTTTMPMSTTSERSTLLTT
  355 VLISPISVMSPSEASTLSTPPGDTSTPLLTSTKAGSFSIPAEVTTIRISITSERSTPLT
  414 TLLVSTTLPTSFPGASIASTPPLDTSTTFTPSTDTASTPTIPVATTISVSVITEGSTPG
  473 TTIFIPSTPVTSSTADVPATTGAVSTPVITSTELNTPSTSSSSTTTSFSTTKEFTTPAM
  532 TTAAPLTYVTMSTAPSTPRTTSRGCTTSASTLSA
  566 TSTPHTSTSITTTETPS
  583 SSTPSFTSSITTTETTS
  600 HSTPSFTSSISTTETTS
  617 HSTPSFTSLITITETTS
  634 HSTPSYTTSITTTE
```

FIG. 3. MUC3 amino acid sequence located upstream of the 17 amino acid repeat region. Translated sequences of pMUC3T9 and pMUC3T10 are shown respectively in panels A and B, and correspond to the cDNAs shown respectively in panels 2C and D. The amino acid sequence in panel C is derived from the cDNA sequence shown in figure 2B. The consensus 59 amino acid sequence is indicated at the top of the figure. Underlined are the amino acids in the 59 amino acid repeats that are different from the consensus sequence. Asterisks mark potential *N*-glycosylation sites.

ASSEASTLSTTPVDTSTPVTTSTQASSSPTTAEGTSMPTSTPSEGSTPLTSMP of the 59 amino acid repeat (figure 3). No potential *N*-glycosylation sites are present within this consensus sequence and the consensus sequence of the 17 amino acid repeat [HSTPSFTSSITTTETTS] (7). However, occasionally potential *N*-glycosylation sites are present within the 59 amino acid repeat region (figure 3) and 17 amino acid repeat (7). All 59 amino acid repeats are very well conserved (60%-85%). It is of note that the 4 repeats located just upstream of the semi-unique region are the least conserved (60%-73%).

DISCUSSION

Here we report the results of construction and three subsequent screenings of a non-unamplified human small intestinal cDNA library in order to clone non-repetitive cysteine-rich MUC3 cDNA sequences. Although from these three subsequent screenings we obtained in total 27 recombinant clones containing MUC3 cDNA sequences, none of these sequences were non-repetitive and cysteine-rich, which are characteristic features of the N- and C-terminal flanking sequences of many mucins. Instead, we report the identification of a novel 59 amino acid repeat located upstream of the earlier identified 17 amino acid repeat of human MUC3. Identification of a second repeat region in a mucin is not without precedent. In MUC2 also two repetitive regions have been identified (6). And similar to MUC2, a non-repetitive semi-unique region is located in between these two repeat regions. However, in contrast to MUC2 this semi-unique region in MUC3 is relatively large (270 amino acids) and is rich in serine-, threonine- and proline-residues. In addition, the consensus sequences of two repetitive regions within MUC3 are very different in length as well as in sequence. Moreover, the 59 amino acid repeat is also distinct from any other mucin repeat, including the repeat identified in rat MUC3 (18,19).

In figure 3, thirteen 59 amino acid repeats are shown, indicating that this repeat region is at least 2.3 kb in length. However, most likely this repeat region comprises at least 5 kb, because one of the non-overlapping MUC3 cDNA clones, pMUC3T4, was 2.7 kb in length and sequence analyses of both ends of the insert revealed the 59 amino acid repeat. In addition, we have cloned 5 non-overlapping partial cDNAs containing the 17 amino acid repeat region that comprises in total about 4 kb in length. On Northern blot MUC3 mRNA has been estimated to be approximately 8 kb in length

(7). This would indicate that the complete MUC3 cDNA consists of repeats. However, the resolution of agarose gels and the fact that MUC3 mRNA on Northern blots usually results in a polydisperse signal does not allow an accurate estimation of mRNA size. However, based on the molecular mass of the endo-H-treated MUC3 protein precursor of 550 kDa (15), we estimate the MUC3 mRNA to be at least about 16.5 kb. Therefore, we conclude that the 59 amino acid repeat region together with the 17 amino acid repeat region encompass at least two-thirds of the full-length MUC3 cDNA.

From a technical point of view, it is interesting to note that thus far all isolated mucin cDNAs have been proven to be partial and frequently quite short (4-12). The MUC3 cDNA fragments reported here have an average length of 1.2 kb. Therefore, the extraordinarily large size of the MUC3 repetitive regions leads to an overrepresentation of these repetitive sequences within the cDNA library which may explain why thus far we did not succeed in cloning the non-repetitive flanking sequences of MUC3.

Very recently, the C-terminal non-repetitive region of rat MUC3 has been cloned (20). Since the non-repetitive, in contrast to the repetitive sequences, in many mucins are usually very well conserved, this finding might be of great help in cloning the C-terminus of its human homologue.

ACKNOWLEDGMENTS

This research was supported by grants from ASTRA Pharmaceuticals Ltd (to BJWvK), foundation 'de Drie Lichten' (to BJWvK), Nutricia (to HAB), and the Netherlands Digestive Diseases Foundation (to AWCE). We are grateful to Dr. H. Vos, Amsterdam, The Netherlands for providing the SIB139 cDNA probe.

REFERENCES

1. Van Klinken, B. J. W., Dekker, J., Büller, H. A., and Einerhand, A. W. C. (1995) *Am. J. Phys.* **269,** G613–G627.
2. Ligtenberg, M. J. L., Vos, H. L., Gennissen, A. M. C., and Hilkens, J. (1990) *J. Biol. Chem.* **265,** 5573–5578.
3. Bobek, L. A., Tsai, H., Biesbrock, R. A., and Levine, M. J. (1993) *J. Biol. Chem.* **268,** 20563–20569.
4. Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lamport, D. T. A., and Kim, Y. S. (1989) *J. Biol. Chem.* **264,** 6480–6487.
5. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E., and Kim, Y. S. (1992) *J. Biol. Chem.* **267,** 21375–21383.
6. Gum, J. R. Jr., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) *J. Biol. Chem.* **269,** 2440–2446.
7. Gum, J. R., Hicks, J. W., Swallow, D. M., Laglace, R. L., Byrd, J. C., Lamport, D. T. A., Siddiki, B., and Kim, Y. S. (1990) *Biochem. Biophys. Res. Comm.* **171,** 407–415.
8. Porchet, N., Ngyen Van Cong, J., Dufosse, J., Audie, J-P., Guyonnet-Duperat, V., Gross, M. S., Denis, C., Degand, P., Bernheim, A., and Aubert, J-P. (1991) *Biochem. Biophys. Res. Comm.* **175,** 414–422.
9. Guyonnet-Duperat, V., Audie, J-P., Deabailleul, V., Laine, A., Buisine, M-P., Galiegue-Zouitina, S., Pigny, P., Degand, P., Aubert, J-P., and Porchet, N. (1995) *Biochem. J.* **305,** 211–219.
10. Dufosse, J., Porchet, N., Audie, J. P., Guyonnet-Duperat, V., Laine, A., Van Seuningen, I., Marrakchi, S., and Aubert, J. P. (1993) *Biochem. J.* **293,** 329–337.
11. Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W-L., Gum, E., Hicks, J. W., Gum, J. R., Byrd, J. C., Siddiki, B., and Kim, Y. S. (1993) *J. Biol. Chem.* **268,** 5879–5885.
12. Shankar, V., Gilmore, M. S., Elkins, R. C., and Sachdev, G. P. (1994) *Biochem. J.* **300,** 295–298.
13. Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., and Petersen, G. M. (1991) *J. Clin. Invest.* **88,** 1005–1013.
14. Tytgat, K. M. A. J., Büller, H. A., Opdam, F. J. M., Kim, Y. S., Einerhand, A. W. C., and Dekker, J. (1994) *Gastroenterology* **107,** 1352–1363.
15. Van Klinken, B. J. W., Oussoren, E., Weenink, J. J., Strous, G. J., Büller, H. A., Dekker, J., and Einerhand, A. W. C. (1995) *Glycoconj. J.* **13,** 757–768.
16. Van Klinken, B. J. W., Dekker, J., Büller, H. A., De Bolòs, C., and Einerhand, A. W. C. (1997) *Am. J. Phys.* **273,** in press.
17. Church, G. M., Gilbert, W. (1984) *Proc. Natl. Acad. Sci.* **81,** 1991–1995.
18. Khatri, I. A., Forstner, G. G., Forstner, J. F. (1993) *Biochem. J.* **294,** 391–399.
19. Gum, J. R., Hicks, J. W., Lagace, R. E., Byrd, J. C., Toribara, N. W., Siddiki, B., Fearney, F. J., Lamport, D. T. A., and Kim, Y. S. (1991) *J. Biol. Chem.* **266,** 22733–22738.
20. Khatri, I. A., Forstner, G. G., Forstner, J. F. (1997) *Biochim. Biophys. Acta* **1326,** 7–11.