Cloning and analysis of human gastric mucin cDNA reveals two types of conserved cysteine-rich domains

Leo W. J. KLOMP,* Linde VAN RENS and Ger J. STROUS†
Laboratory of Cell Biology, AZU H02.314, University of Utrecht, Heidelberglaan 100, AZU rm H02.314, 3584 CX Utrecht, The Netherlands

Human gastric mucin was isolated by successive CsCl-gradient ultracentrifugation in the presence of guanidinium hydrochloride to prevent degradation of the polypeptide moieties of the molecules. The amino acid sequence of a tryptic fragment of this molecule was identical to that of a tryptic fragment of tracheobronchial mucin. An oligonucleotide based on this sequence hybridized specifically to human stomach mRNA and was subsequently used to screen a human stomach λZAPII cDNA library. The largest of 10 positive clones encoded 850 amino acid residues, including the tryptic fragment, with high amounts of threonine, serine and proline residues. Interestingly, cysteine accounted for almost 8 % of the amino acid residues. The 3′ part of the sequence was very similar but not identical to the 3′ region of human tracheobronchial cDNA. No tandem repeated sequences were present and the deduced polypeptide sequence contained two potential N-linked glycosylation sites. Four cysteine-rich clusters were detected, one of which was apparently homologous to the D-domains present in other mucins and in von Willebrand factor. The arrangement of the cysteines in three other cysteine-rich clusters was conserved in the human gastric mucin cDNA in a similar fashion as in two domains in the *MUC2* gene product. The cysteine-rich domains were separated by short stretches of non-repetitive amino acid residues with a very high content of threonine and serine residues. These data suggest that the encoded polypeptide of this clone may be involved in disulphide-bond-mediated oligomerization of the mucin, and provide new insights into the molecular organization of mammalian apomucins.

INTRODUCTION

The first line of defence against noxious agents, bacteria and mechanical forces of the stomach wall and other epithelia of the gastrointestinal and respiratory tracts is the mucus layer. This is a highly viscous gel composed mainly of water and mucus glycoproteins (mucins). Mucins constitute a family of high-molecular-mass glycoproteins secreted by specialized cells in the epithelium [1–3]. Isolated mucins can account entirely for the viscous aspects of mucus [4]. Therefore, understanding the complex structure of these macromolecules is the key to comprehension of the protective function of mucus. An important and characteristic feature of mucins is the high proportion (> 50 %) of O-linked oligosaccharide chains. These chains may variably terminate in sialic acid or sulphate residues, rendering the mucins very hydrophilic and negatively charged [1,2]. Moreover, the physicochemical properties of mucus are dependent on a disulphide-bond-mediated polymeric configuration of the mucins, as the viscosity is greatly affected after addition of reducing agents [4,5]. Isolated mucins from different sources appear as filamentous thread-like homo-oligomeric structures, with lengths up to 10 μm, which can be chemically reduced to smaller subunits [6–10].

Previously we described the structure of prominent gastric mucins isolated from rat and human stomach tissue [10,11]. Both mucins (designated RGM and HGM respectively) contained a large proteinase-resistant domain, containing most of the O-linked oligosaccharide chains. In addition, proteinase-sensitive domains could be discerned, containing all of the cysteine residues [10,11]. Polyclonal antisera raised against these mucins were directed towards the polypeptide moieties and displayed interspecies cross-reactivity. The early biosynthetic precursors of RGM and HGM were identified as 300 kDa and 500 kDa polypeptides respectively. They contained N-linked glycans and constituted disulphide-linked oligomers shortly after translocation into the endoplasmic reticulum (ER) [12–14].

Additional information on the structure of mucin polypeptide backbones was deduced from cDNA cloning and sequencing. At least seven different human mucin genes have been identified thus far, one of which (*MUC1*) encodes a carcinoma-associated membrane-bound mucin [15]. Most of the other identified cDNAs (*MUC2* to *MUC6*) are likely to code for secreted gel-forming mucins, expressed in different epithelia [16–24]. From these studies a general model of the molecular structure of mucins has emerged, in accordance with the biochemical data described above. Each of these mucins contains a variable number of unique tandemly repeated sequences, ranging from eight (MUC5c [22]) to 169 (MUC6 [24]) amino acid residues. They are composed mainly of threonine and serine residues and comprise most of the O-linked glycan attachment sites. In the MUC2 gene product, the major mucin of the human intestine [25], these tandem repeats are flanked by cysteine-rich domains of approx. 350 amino acid residues. These cysteine-rich domains are homologous to the D-domains of human von Willebrand factor (vWF) and have been implicated in oligomerization of the MUC2 protein [18,19]. These domains contain considerably fewer potential O-glycosylation sites. Cysteine-rich domains have also

Abbreviations used: BSM, bovine submaxillary mucin; FIM, frog integumentary mucin; HGM, human gastric mucin; ER, endoplasmic reticulum; PSM, porcine submaxillary mucin; RGM, rat gastric mucin; SSC, standard saline citrate; TFA, trifluoroacetic acid; TBM, tracheobronchial mucin; vWF, von Willebrand factor.

* Present address: Laboratory for Immunology and Rheumatology, Washington University, Childrens Hospital, 1 Children's Place, St. Louis, MO, U.S.A.
† To whom correspondence should be addressed.

The nucleotide sequence data reported are available from the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under accession no. X81649.

been reported in rat mucin-like protein (the rat homologue of MUC2), porcine and bovine submaxillary mucins (PSM, BSM), human tracheobronchial mucin (TBM) and *Xenopus laevis* integumentary mucins (FIMs), and can therefore be regarded as typical for gel-forming mucins [23,26,27–31].

To elucidate the molecular structure of the HGM polypeptide and backbone and to better understand the mechanism of oligomerization, we isolated unique cDNA clones corresponding to a non-tandem-repeat region, by exploiting an oligonucleotide probe corresponding to a proteinase-sensitive domain in HGM. The isolated clones appeared to have sequence identity to TBM cDNA [23,32] and encoded several cysteine-rich domains.

MATERIALS AND METHODS

Cell line and human tissue

The Hep-G2 cell line was propagated as described [33]. Stomach tissue was obtained from oesophagus-carcinoma patients undergoing partial stomach resection in the Academic Hospital of Utrecht. Only stomach tissue that was healthy by macroscopic criteria was included in this study.

Materials

Restriction nucleases, modifying enzymes, trypsin and standard molecular biology reagents (analytical grade) were from Boehringer (Mannheim, Germany). All ^{32}P and ^{35}S radioisotopes were from Amersham. The Superscript cDNA synthesis system and Sequenase 2.0 were purchased from BRL; λZAPII, XL1blue and the Gigapack Gold packaging kit were from Stratagene (La Jolla, CA, U.S.A.); TA-cloning kit was from Invitrogen (San Diego, CA, U.S.A.); and the Taq dye cycle sequencing kit was from Applied Biosystems. DNA probes were labelled using the oligonucleotide labelling kit from Pharmacia.

Oligonucleotides were synthesized on an Applied Biosystems 392 DNA synthesizer. PCR was performed in a Perkin–Elmer 480 thermocycler using AmpliTaq polymerase (Perkin–Elmer, Norwalk, CT. U.S.A.).

Isolation of gastric mucin and generation of tryptic fragments

HGM was isolated from pooled fundic mucosae of four oesophagus-carcinoma patients exactly as described [11,34]. The material was analysed by SDS/PAGE and silver staining and no contaminating proteins were found. The mucin was reduced and alkylated using iodoacetamide, dialysed against water and lyophilized. A 1.5 mg sample of this preparation was redissolved in 100 mM NH_4HCO_3, trypsin was added to 0.04 M and the digestion was allowed to proceed at 37 °C for 14 h. A trypsin self-digestion was also initiated to avoid sequencing possible fragments of trypsin itself. The digest was acidified to 0.1% with trifluoroacetic acid (TFA) and a 50 μl aliquot was chromatographed on a Brownlee RP-300 (C_8-RPLC) column (2.1 mm × 30 mm), which was eluted using a gradient of acetonitrile in 0.1% TFA. A total of 11 peaks were identified, and designated HGM-t1 to HGM-t11. HGM-t11, a sharp peak eluting at 28% acetonitrile, was the most abundant fragment which by retention time did not appear to represent fragments of trypsin itself. This fragment was subjected to Edman degradation by an automated Applied Biosystems 470A sequencer.

Preparation and screening of a human stomach cDNA library

For the isolation of mRNA, gastric tissue was frozen in liquid N_2 immediately following removal from the patient. A 1 g portion of tissue was ground to a fine powder; after dissolution in guanidinium isothiocyanate, RNA was isolated as described [35]. mRNA was subsequently isolated according to standard protocols [36,37]. A 5 μg sample of poly(A)$^+$ RNA was used to generate cDNA with the Superscript system, using random hexameric primers, according to the manufacturer's manual. A cDNA fraction containing material of > 1 kbp was used to ligate into λZAPII, and the resulting library (6.6×10^5 recombinants) was amplified once and plated on to *Escherichia coli* XL1blue. The library was plated (1×10^5 recombinants) and screened in duplicate with a 66-mer oligonucleotide corresponding to HGM-t11 (see Table 1). Filters were prehybridized in 6 × SSC, 0.1% SDS, 20% formamide, 0.05 M sodium phosphate, pH 7.0, 5 × Denhardt's, 0.05% sodium pyrophosphate, 10% dextran sulphate, 100 μg/ml freshly denatured salmon sperm DNA and 50 μg/ml yeast tRNA at 42 °C for 4 h. A 20 pmol sample of the HGM-t11 oligonucleotide was end-labelled with ^{32}P to 10^8 c.p.m./μg using T4-polynucleotide kinase and the probe was added to the hybridization solution at 1×10^6 c.p.m./ml. After hybridization for 16 h, the filters were washed and exposed to Kodak X-AR films overnight after a final wash at 55 °C in 0.1 × SSC, 0.1% SDS and 0.05% sodium pyrophosphate. Positives were re-screened until free from contaminating plaques and were excision-rescued.

Northern and Southern blot analysis

RNA samples (20 μg) were separated on 0.9% agarose gels containing 2.2 M formaldehyde [37]. Gels were treated with 50 mM NaOH for 20 min and neutralized by extensive washing in 40 mM Mops, 5 mM sodium acetate and 10 mM NaEDTA, pH 7.0. Capillary transfer of the RNA to nitrocellulose was preceded by a 30 min wash with 10 × SSC and performed overnight using 10 × SSC rather than 20 × SSC. Alternatively, multiple human tissue poly(A)$^+$ RNA blots were purchased from Clontech. Southern blot analysis of cDNA clones was performed by agarose-gel separation of 1 μg of the plasmid digested with the appropriate restriction enzymes followed by standard blotting procedures [37]. Hybridization and washing of the Southern and Northern blots with the HGM-t11 oligonucleotide were performed under similar conditions as used to screen the library. The complete inserts of isolated HGM clones or different fragments thereof (200 bp *Pst*I fragment of clones HGM-1 and HGM-5 or the 5′ 750 bp *Pst*I fragment of HGM-1) were used as HGM cDNA probes. Probes were isolated by digestion with the appropriate restriction enzymes and separation on agarose gels, labelled with [α-^{32}P]dCTP and used at 1×10^6 c.p.m./ml. These probes were hybridized similarly, but the filters were subjected to a final stringency wash of 30 min at 65 °C in 0.1 × SSC, 0.1% SDS and 0.05% sodium pyrophosphate.

cDNA sequencing and PCR

Restriction maps of the isolated clones were made and overlapping subclones were generated in pUC19. Template DNA for sequencing was prepared as described previously [38]. Sequencing reactions were carried out using the Taq dye primer cycle sequencing kit with fluorescently labelled −21M13 forward and reverse primers (Applied Biosystems) in a Perkin–Elmer 9600 thermocycler. Sequence reactions were analysed on an Applied Biosystems model 373A sequencer. Ambiguous fragments were resequenced by conventional methods using gene-specific primers (GSP-1–4) (Table 1) and Sequenase 2.0 (USB, Cleveland, OH, U.S.A.). Sequences were analysed using the PC-gene Intelli-Genetics software.

RESULTS

HGM amino acid sequence information

No sequences were obtained when the purified HGM was subjected to Edman degradation sequencing, probably indicating that the mucin was N-terminally blocked. Since it was previously shown that small HGM fragments could be generated by trypsin digestion [11], tryptic fragments were sequenced. The sequence of the most abundant fragment, HGM-t11, could be unambiguously determined as: Phe-Asp-Val-Asp-Phe-Pro-Ser-Pro-Gly-Pro-His-Gly-Gly-Asp-Lys-Glu-Thr-Tyr-Asn-Asn-Ile-Ile. Surprisingly, this 22-amino-acid sequence exactly matched the sequence of the TBM:TR-3a tryptic fragment of human TBM [32]. Compared with that sequence, an extra isoleucine and arginine were detected at the C-terminus of HGM-t11. The assignment of an N-terminal tryptophan residue was ambiguous in our sequence.

A '66-guessmer' oligonucleotide (Table 1) corresponding to the sequence of HGM-t11 was synthesized and used to probe Northern blots to investigate the feasibility of using the oligonucleotide in cDNA cloning studies. The sequence was deduced using the recommended codons when designing oligonucleotide probes [37]. A high-molecular-mass transcript, estimated to be > 10 kb, was detected in human stomach total RNA (Figure 1) as well as in poly(A)$^+$ RNA (not shown). Only upon overexposure of the blots was a weak signal noticed in gall bladder RNA, but the oligonucleotide did not hybridize to RNA isolated from Hep-G2 cells (results not shown). Under similar stringency conditions this transcript was not detected in rat stomach RNA or human lung RNA (results not shown). The length of the mRNA and the specific hybridization pattern suggested that the HGM-t11 oligonucleotide indeed was suitable for the isolation of HGM cDNA clones.

Molecular cloning of HGM cDNA and analysis of expression

Next, a human stomach cDNA library was constructed in λZAPII. Since the location of HGM-t11 in the protein (N-terminal or C-terminal) was unknown, and since the HGM transcript was larger than 10 kb, precautions were taken to generate a library containing both 5′ and 3′ sequences of the expressed genes. During our studies the importance of using high-quality mRNA became apparent, as 5′ sequences may easily be under-represented after poly(A)$^+$ selection when degraded RNA is used. The first-strand cDNA synthesis was primed with random hexamers rather than poly(dT). The presence of a 377 bp 5′ fragment of the gastric H^+/K^+-ATPase (corresponding to nucleotides 22–399 [39]) was tested at each step of generation of the library. This was done by subjecting small aliquots of the nascent library to PCR analysis with gene-specific primers HKATPaseNf and HKATPaseNr (Table 1). In all PCR analyses, including analysis of the ultimate library, the amplified 5′ H^+/K^+-ATPase cDNA fragment appeared with comparable intensity to 3′ H^+/K^+-ATPase cDNA fragments (nucleotides 2312–3078 in [39] amplified by HKATPaseCf and HKATPaseCr; Table 1) or 3′ pepsinogen C cDNA (primers PEPC-Cf and PEPC-Cr; nucleotides 270–370 [40]). In addition a 342 bp MUC6 fragment [24] could be PCR-cloned from the library using the MUC6-specific primers described in Table 1. This suggested that the library was appropriate to isolate HGM cDNA clones with the HGM-t11 oligonucleotide, including the possibility of 5′ end location.

After screening 100000 clones, 13 clones appeared positive in duplicate, and 10 of these were further purified by rescreening. After *in vivo* excision of the Bluescript plasmid, the inserts were released with *Not*I and digested with *Pst*I and *Sac*I. The restriction fragments were subsequently tested on Southern blot for hybridization with the HGM-t11 probe, revealing nine clones with positive inserts (results not shown). These clones were designated as HGM-1 to HGM-9.

The clone with the largest insert, HGM-1, was selected for further analysis. To validate the identity of this clone as a candidate HGM cDNA, we labelled the complete insert to test expression of this cDNA in high-stringency Northern-blotting experiments. On similar blots a signal was obtained which was indistinguishable from that of the HGM-t11 probe. Again, a > 10 kb transcript was detected in human stomach (Figure 1). The figure shows variable polydispersity of the signal among different experiments. This might be explained by variable degradation of

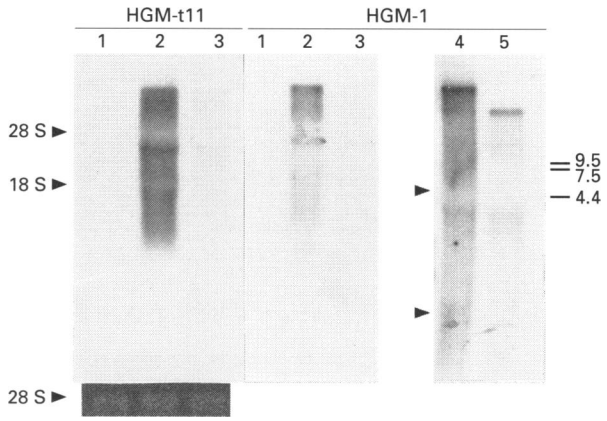

Figure 1 Northern blot analysis of HGM expression

Northern blots containing 20 μg of Hep-G2 RNA (lanes 1), human stomach RNA (lanes 2 and 4), human gall-bladder RNA (lanes 3) or rat stomach RNA (lane 5) were probed with the HGM-t11 oligonucleotide (HGM-t11) or with the HGM-1 insert (HGM-1) and exposed overnight. Migration of RNA is indicated on the left. The ethidium- bromide-stained 28 S rRNA is shown to verify equal loading of each lane. The numbers on the right indicate the migration of RNA markers for lanes 4 and 5 only.

Table 1 Primers and oligonucleotide used for PCR, sequencing and hybridization

Nucleotides added to the primers to generate *Bam*HI restriction sites are underlined

Name	Sequence	Ref.
HGM-t11	5′-TTTGATGTGGACTTCCCCTCCCCTGGCCCCCAT-GGTGGTGACAAGGAGACCTACAACAACATCATC-3′	
MUC6f	5′-TGGGATCCTCTTCCACACGTCCCATGACGGCA-3′	24
MUC6r	5′-TGGGATCCATGTGAGTGGAGGGATGTAGAGGT-3′	
PEP-Cf	5′-TGGGATCCCAGCAGTACATGAGTGCTCTTC-TGCAGGCCAC-3′	40
PEP-Cr	5′-TGGGATCCAGTGGCAAAGCCTACTCTGT-TGTTGCCCAAGT-3′	
HKATPaseNf	5′-TGGGATCCGAGCTCTACTCGGTGGAGCTGG-3′	39
HKATPaseNr	5′-TGGGATCCATTGTCGTCGGTGGTGAGGTCC-3′	
HKATPaseCf	5′-TGGGATCCGTCGACTGATCTTCGACAACCT-3′	
HKATPaseCr	5′-TGGGATCCTCCCTGGGCAACAGCGAACTCC-3′	
GSP-1	5′-CAGAAGCAGTGCAGCATCC-3′	
GSP-2	5′-CACACACTGCATCTTGTCC-3′	
GSP-3	5′-CAAAGCTGAGGCCTGTGTC-3′	
GSP-4	5′-GACATCCATCCATGGCGAC-3′	

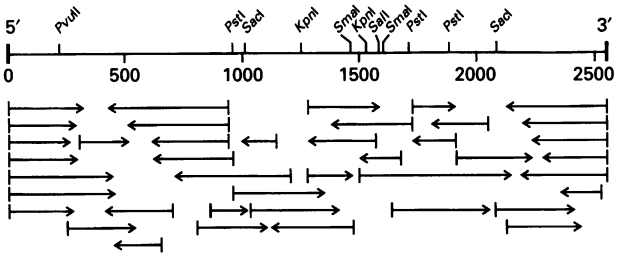

Figure 2 Restriction map and sequencing strategy of clone HGM-1

Restriction sites utilized for subcloning are shown on the top. The arrows indicate the length and the direction of the different sequences obtained.

the RNA samples from different patients. After stringent washing the signal was absent from Hep-G2 RNA and gall-bladder RNA, but after overexposure of the blot a faint transcript was detected in human gall bladder (results not shown). Similar transcripts were detected using clones HGM-5, HGM-6 or restriction fragments of HGM-1 as probes. The tissue expression of the mRNA was further tested by hybridization of multiple tissue Northern blots. No expression was observed in non-mucin-producing tissues (e.g. heart, brain and muscle), nor did we find the transcripts in the lung, pancreas, colon or the small intestine, even after overexposure of the blots (results not shown). Interestingly, a distinct transcript was detected in the rat stomach (Figure 1), suggesting sequence similarity between the HGM and RGM genes, although the rat mRNA appeared clearly smaller than its human counterpart. These data suggest that HGM-1 represents a valid gastric mucin cDNA clone which is expressed in human and rat stomach.

Sequence determination and analysis of HGM-1

Clone HGM-1 was sequenced completely according to the strategy described in Figure 2. Due to an extremely high proportion of GC residues (63%), many sequence problems

```
⇔ domain I
GGAC ACC GAC ATT GGC CTG GTG CTG CTG TGG GAC AAG AAG ACC AGC ATC TTC ATC     55
 Asp Thr Asp Ile Gly Leu Val Leu Leu Trp Asp Lys Lys Thr Ser Ile Phe Ile     18

AAC CTC AGC CCC GAG TTC AAG GGC AGG GTC TGC GGC CTG TGT GGG AAC TTC GAC    109
Asn Leu Ser Pro Glu Phe Lys Gly Arg Val Cys Gly Leu Cys Gly Asn Phe Asp     36
 *
GAC ATC GCC GTT AAT GAC TTT GCC ACG CGG AGC CGG TCT GTG GTG GGG GAC GTG    163
Asp Ile Ala Val Asn Asp Phe Ala Thr Arg Ser Arg Ser Val Val Gly Asp Val     54

CTG GAG TTT GGG AAC AGC TGG AAG CTC TCC CCC TCC TGC CCA GAT GCC CTG GCG    217
Leu Glu Phe Gly Asn Ser Trp Lys Leu Ser Pro Ser Cys Pro Asp Ala Leu Ala     72

CCC AAG GAC CCC TGC ACG GCC AAC CCC TTC CGC AAG TCC TGG GCC CAG AAG CAG    271
Pro Lys Asp Pro Cys Thr Ala Asn Pro Phe Arg Lys Ser Trp Ala Gln Lys Gln     90

TGC AGC ATC CTC CAC GGC CCC ACC TTC GCC GCC TGC CAC GCA CAC GTG GAG CCG    325
Cys Ser Ile Leu His Gly Pro Thr Phe Ala Ala Cys His Ala His Val Glu Pro    108

GCC AGG TAC TAC GAG GCC TGC GTG AAC GAC GCG TGC GCC TGC GAC TCC GGG GGT    379
Ala Arg Tyr Tyr Glu Ala Cys Val Asn Asp Ala Cys Ala Cys Asp Ser Gly Gly    126

GAC TGC GAG TGC TTC TGC ACG GCT GTG GCC CGC TAC GCC CAG GCC TGC CAT GAA    433
Asp Cys Glu Cys Phe Cys Thr Ala Val Ala Arg Tyr Ala Gln Ala Cys His Glu    144

GTA GGC ACC TGT GTG TGT CTG CGG ACC CCA AGC ATC TGC CCT CTG TTC TGC GAC    487
Val Gly Thr Cys Val Cys Leu Arg Thr Pro Ser Ile Cys Pro Leu Phe Cys Asp    162

TAC TAC AAC CCC GAA GGC CAG TGC GAG TGG CAC TAC CAG CCC TGC GGG GTG CCC    541
Tyr Tyr Asn Pro Glu Gly Gln Cys Glu Trp His Tyr Gln Pro Cys Gly Val Pro    180

TGC CTG CGC ACC TGC CGG AAC CCC CGT GGA GAC TGC CTG CGG GAC GTC CGG GGC    595
Cys Leu Arg Thr Cys Arg Asn Pro Arg Gly Asp Cys Leu Arg Asp Val Arg Gly    198

CTG GAA GGC TGC TAC CCC AAG TGC CCA CCA GAG GCT CCC ATC TTT GAT GAG GAC    649
Leu Glu Gly Cys Tyr Pro Lys Cys Pro Pro Glu Ala Pro Ile Phe Asp Glu Asp    216

AAG ATG CAG TGT GTG GCC ACC TGC CCA ACC CCG CCT CTG CCA CCA CGG TGC CAC    703
Lys Met Gln Cys Val Ala Thr Cys Pro Thr Pro Pro Leu Pro Pro Arg Cys His    234

GTC CAT GGG AAG TCC TAC CGG CCA GGT GCA GTG GTG CCC TCG GAC AAG AAC TGC    757
Val His Gly Lys Ser Tyr Arg Pro Gly Ala Val Val Pro Ser Asp Lys Asn Cys    252

CAG TCC TGC CTT TGT ACG GAG CGC GGC GTG GAG TGC ACC TAC AAA GCT GAG GCC    811
Gln Ser Cys Leu Cys Thr Glu Arg Gly Val Glu Cys Thr Tyr Lys Ala Glu Ala    270

TGT GTC TGC ACC TAC AAT GGA CAG CGC TTC CAC CCA GGG GAC GTC ATC TAC CAC    865
Cys Val Cys Thr Tyr Asn Gly Gln Arg Phe His Pro Gly Asp Val Ile Tyr His    288

ACG ACG GAT GGC ACG GGT GGC TGC ATC TCC GCC CGC TGC GGG GCC AAC GGC ACC    919
Thr Thr Asp Gly Thr Gly Gly Cys Ile Ser Ala Arg Cys Gly Ala Asn Gly Thr    306
                                                    *
                    domain I ⇔ ⇔ domain II
ATT GAG AGG AGG GTC TAC CCC TGC AGC CCC ACC ACC CCT GTC CCC CCA ACC ACC    973
Ile Glu Arg Arg Val Tyr Pro Cys Ser Pro Thr Thr Pro Val Pro Pro Thr Thr    324

TTC TCC TTC TCC ACA CCC CCG CTT GTC GTG AGC TCC ACG CAC ACC CCC AGC AAT   1027
Phe Ser Phe Ser Thr Pro Pro Leu Val Val Ser Ser Thr His Thr Pro Ser Asn    342

GGC CCA AGC AGC GCG CAC ACA GGC CCT CCG AGC AGC GCC TGG CCC ACC ACA GCA   1081
Gly Pro Ser Ser Ala His Thr Gly Pro Pro Ser Ser Ala Trp Pro Thr Thr Ala    360
                                                    domain II ⇔
GGC ACT TCT CCC AGG ACG AGG CTG CCC ACA GCC TCT GCC TCG CTG CCG CCG GTC   1135
Gly Thr Ser Pro Arg Thr Arg Leu Pro Thr Ala Ser Ala Ser Leu Pro Pro Val    378

⇔ domain III
TGT GGG GAA AAG TGC CTG TGG TCG CCA TGG ATG GAT GTC AGC CGC CCT GGA CGG   1189
Cys Gly Glu Lys Cys Leu Trp Ser Pro Trp Met Asp Val Ser Arg Pro Gly Arg    396

GGC ACG GAC AGC GGT GAC TTC GAC ACA CTG GAG AAC CTC CGC GCC CAT GGG TAC   1243
Gly Thr Asp Ser Gly Asp Phe Asp Thr Leu Glu Asn Leu Arg Ala His Gly Tyr    414

CGG GTG TGC GAA TCA CCC AGG TCG GTG GAG TGC CGA GCC GAG GAC GCC CCC GGA   1297
Arg Val Cys Glu Ser Pro Arg Ser Val Glu Cys Arg Ala Glu Asp Ala Pro Gly    432

GTG CCG CTC CGA GCC CTG GGG CAG CGT GTG CAG TGC AGC CCG GAT GTG GGG CTG   1351
Val Pro Leu Arg Ala Leu Gly Gln Arg Val Gln Cys Ser Pro Asp Val Gly Leu    450

ACC TGT CGT AAC AGG GAG CAG GCA TCG GGG CTC TGC TAC AAC TAC CAG ATC AGG   1405
Thr Cys Arg Asn Arg Glu Gln Ala Ser Gly Leu Cys Tyr Asn Tyr Gln Ile Arg    468
                        domain III ⇔ ⇔ domain IV
GTC CAG TGC TGC ACG CCC CTA GCC TGC TCC ACC TCT AGC AGT CCA GCC CAG ACC   1459
Val Gln Cys Cys Thr Pro Leu Ala Cys Ser Thr Ser Ser Ser Pro Ala Gln Thr    486

ACT CCT CCA ACT ACC TCC AAG ACC ACT GAA ACC CGG GCC TCA GGC TCC TCA GCT   1513
Thr Pro Pro Thr Thr Ser Lys Thr Thr Glu Thr Arg Ala Ser Gly Ser Ser Ala    504

CCC AGC AGC ACA CCT GGC ACC GTG TCT CTC TCT ACA GCC AGG ACG ACA CCT GCC   1567
Pro Ser Ser Thr Pro Gly Thr Val Ser Leu Ser Thr Ala Arg Thr Thr Pro Ala    522

CCA GGT ACC GCT ACC TCT GTC AAA AAA ACT TTC TCA ACT CCC AGC CCT CCG CCA   1621
Pro Gly Thr Ala Thr Ser Val Lys Lys Thr Phe Ser Thr Pro Ser Pro Pro Pro    540

GTG CCG GCA ACA TCA ACA TCA TCC ATG TCG ACC ACG GCC CCG GGG ACC TCT GTG   1675
Val Pro Ala Thr Ser Thr Ser Ser Met Ser Thr Thr Ala Pro Gly Thr Ser Val    558
                                                    domain IV ⇔ ⇔ domain V
GTC TCC AGC AAG CCC ACC CCC ACG GAG CCC AGC ACA TCC TCC TGC CTG CAG GAG   1729
Val Ser Ser Lys Pro Thr Pro Thr Glu Pro Ser Thr Ser Ser Cys Leu Gln Glu    576

CTT TGC ACC TGG ACC GAG TGG ATC GAT GGC AGC TAC CCT GCT CCT GGA ATA AAT   1783
Leu Cys Thr Trp Thr Glu Trp Ile Asp Gly Ser Tyr Pro Ala Pro Gly Ile Asn    594

GGT GGA GAT TTT GAC ACA TTT CAA AAT TTG AGA GAC GAA GGA TAC ACA TTC TGT   1837
Gly Gly Asp Phe Asp Thr Phe Gln Asn Leu Arg Asp Glu Gly Tyr Thr Phe Cys    612

GAA AGT CCT CGA AGC GTG CAG TGC CGG GCA GAG AGC TTC CCC AAC ACG CCG CTG   1891
Glu Ser Pro Arg Ser Val Gln Cys Arg Ala Glu Ser Phe Pro Asn Thr Pro Leu    630

GCA GAC CTG GGG CAG GAC GTC ATC TGC AGC CAC ACA GAG GGG CTG ATT TGC CTG   1945
Ala Asp Leu Gly Gln Asp Val Ile Cys Ser His Thr Glu Gly Leu Ile Cys Leu    648

AAC AAG AAC CAG CTC CCA CCC ATC TGC TAC AAC TAT GAG ATC CGC ATC CAG TGT   1999
Asn Lys Asn Gln Leu Pro Pro Ile Cys Tyr Asn Tyr Glu Ile Arg Ile Gln Cys    666
                        domain V ⇔ ⇔ domain VI
TGC GAG ACG GTG AAC GTG TGC AGA GAC ATC ACC AGA CCG CCA AAG ACC GTC GCA   2053
Cys Glu Thr Val Asn Val Cys Arg Asp Ile Thr Arg Pro Pro Lys Thr Val Ala    684

ACG ACA CGG CCG ACT CCA CAT CCA ACC GGA GCT CAG ACC CAG ACC ACC TTC ACC   2107
Thr Thr Arg Pro Thr Pro His Pro Thr Gly Ala Gln Thr Gln Thr Thr Phe Thr    702

ACA CAC ATG CCC TCG GCC TCC ACA GAG CAA CCC ACG GCA ACC TCC AGG GGT GGG   2161
Thr His Met Pro Ser Ala Ser Thr Glu Gln Pro Thr Ala Thr Ser Arg Gly Gly    720
                                                                domain VI ⇔
CCC ACA GCA ACC AGC GTC ACA CAG GGC ACC CAC ACC ACA CCA GTC ACC AGA AAC   2215
Pro Thr Ala Thr Ser Val Thr Gln Gly Thr His Thr Thr Pro Val Thr Arg Asn    738

⇔ domain VII
TGT CAT CCC CGG TGC ACC TGG ACA AAG TGG TTC GAC GTG GAC TTC CCG TCC CCC   2269
Cys His Pro Arg Cys Thr Trp Thr Lys Trp Phe Asp Val Asp Phe Pro Ser Pro    756

GGA CCC CAT GGT GGA GAC AAG GAA ACC TAC AAC AAC ATC ATC AGG AGT GGG GAA   2323
Gly Pro His Gly Gly Asp Lys Glu Thr Tyr Asn Asn Ile Ile Arg Ser Gly Glu    774

AAA ATC TGC CGC CGA CCT GAG GAG ATC ACC AGG GTC CAG TGC CGA GCC AAG AGC   2377
Lys Ile Cys Arg Arg Pro Glu Glu Ile Thr Arg Val Gln Cys Arg Ala Lys Ser    792

CAC CCA GAG GTG AGC ATC GAA CAC CTG GGC CAG GTG GTG CAG TGC AGC CGG GAA   2431
His Pro Glu Val Ser Ile Glu His Leu Gly Gln Val Val Gln Cys Ser Arg Glu    810

GAG GGC CTG GTG TGC CGG AAC CAG GAC CAG CAG GGA CCC TTC AAG ATG TGC CTC   2567
Glu Gly Leu Val Cys Arg Asn Gln Asp Gln Gln Gly Pro Phe Lys Met Cys Leu    828
                                                        domain VII ⇔
AAC TAC GAG GTG CGT GTG CTC TGC TGC GAG ACC CCC AGA GGC TGC CAC ATG ACC   2539
Asn Tyr Glu Val Arg Val Leu Cys Cys Glu Thr Pro Arg Gly Cys His Met Thr    846

TCC ACA CCT GGC                                                           2551
Ser Thr Pro Gly                                                            850
```

Figure 3 Nucleotide and deduced amino acid sequence of HGM-1

The nucleotide sequence of HGM-1 was determined as described in Figure 2. The deduced amino acids are shown under the nucleotide sequence. The sequence of the HGM-t11 tryptic fragment is underlined. Two potential N-glycosylation sites are indicated by asterisks. Cysteine residues are underlined. Seven domains (see text) are indicated by double arrows.

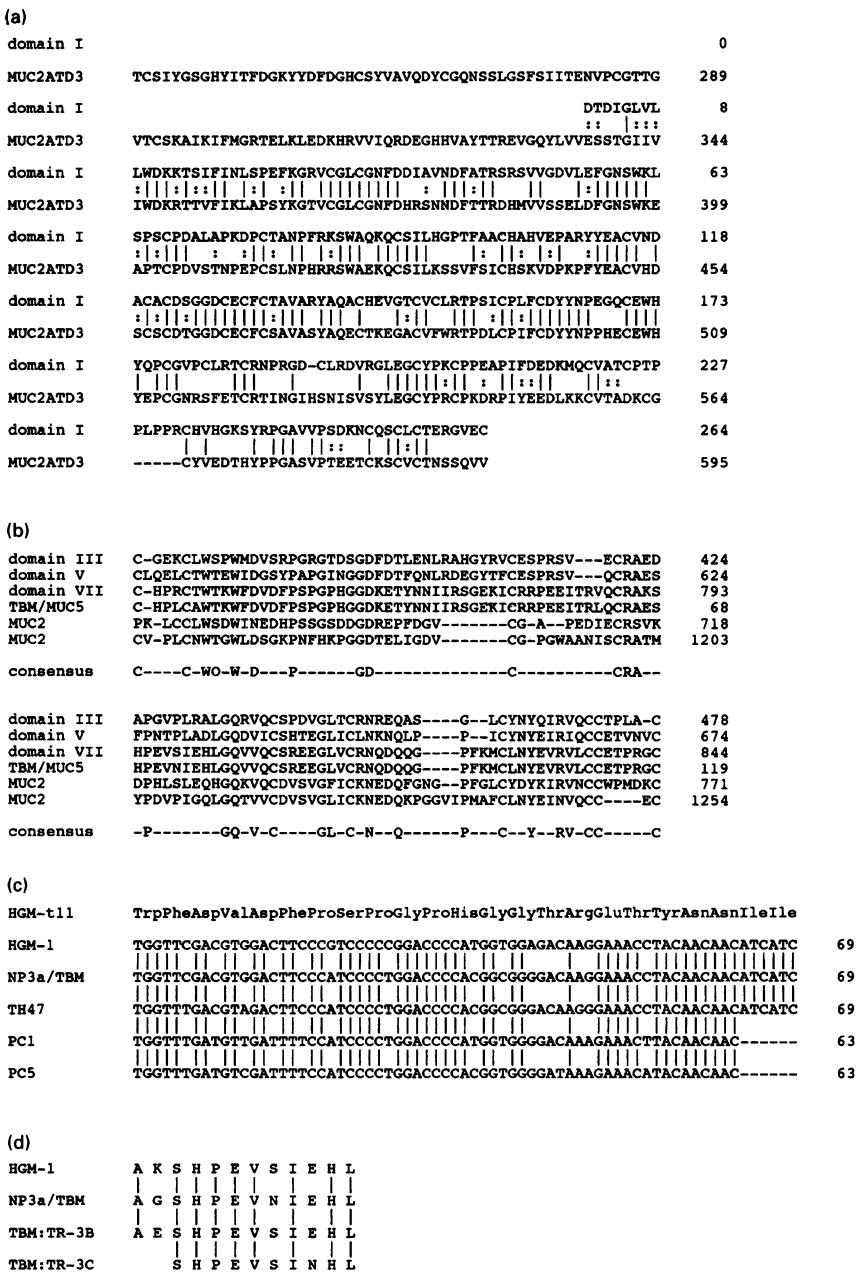

Figure 4 Sequence similarity between MUC2, TBM cDNA and HGM-1

(**a**) Alignment using the PC gene Align program of the third D-domain of MUC2 (MUC2ATD3 [18,19]) with HGM-1 domain I. Identical residues (single-letter code) are indicated with |; conservative substitutions are indicated with :. (**b**) Amino acid residues (single-letter code) of three cysteine-rich domains of HGM-1 were aligned with two analogous domains in a region of MUC2 N-terminal to the tandem repeats [18] and with an analogous domain in TBM cDNA [23] using the Clustall program. The enumeration of the sequences in the mentioned articles was used. The consensus was established when five out of six residues were identical. O indicates serine or threonine. (**c**) Alignment of cDNA sequences found to encode the HGM-t11/TBM:TR-3a tryptic fragment [22,23]. TBM represents the sequence of clone NP3a, which was cloned from a nasal polyp cDNA library [23]. PC1 and PC5 represent clones amplified from nasal polyp RNA by reverse transcriptase PCR [23] and TH47 represents the sequence of clone TH47 in [22]. Amino acid residues are in three-letter code. | indicates identity of all sequences. Note that none of the sequences are completely identical. (**d**) Representation of tryptic fragments TBM:TR-3b and HGM-t11/TBM:TR-3c [31] deduced amino acid residues of clones HGM-1 (Figure 3) and NP3a (TBM) [23] similar to these polypeptides. Amino acid residues are in one-letter code.

were encountered. HGM-1 subclones were therefore sequenced several times using fluorescent-primer-based automatic sequencing before a reliable sequence was obtained. Ambiguous parts were resequenced twice, using conventional techniques.

Clone HGM-1 contained 2551 bp, encoding an open reading frame of 850 amino acid residues (Figure 3). The HGM-t11 tryptic fragment was identified and facilitated orientation of the appropriate reading frame. Nucleotides 2177–2536 had considerable sequence identity (91.7 % at the nucleotide level; 86.7 % at the amino acid level) with the extreme 5′ 360 bp of clone NP3a, a recently described cDNA encoding the C-terminus of human TBM [23]. In addition to the HGM-t11/TBM:TR-3a tryptic fragment of HGM and TBM [32], we identified in this region a stretch of amino acids homologous to the TBM:TR-3b

Table 2 Amino acid composition of the seven HGM-1 subdomains

Amino acid compositions were determined from the deduced cDNA sequence shown in Figure 3 and are depicted as percentages. HGM-1 indicates the composition of the complete cDNA clone.

	Amino acid composition by domain (%)							
Residue	HGM-1	I	II	III	IV	V	VI	VII
Ala	6.1	7.6	7.8	6.0	8.4	2.9	7.6	0.9
Arg	6.2	6.0	3.1	11.1	2.1	3.9	7.6	9.5
Asn	3.1	3.5	1.5	3.0	0.0	6.9	1.5	3.8
Asp	4.3	6.3	0.0	6.0	0.0	5.9	1.5	3.8
Cys	7.7	11.4	0.0	10.1	0.0	9.9	0.0	9.5
Gln	3.5	2.5	0.0	5.0	1.0	5.9	6.1	5.7
Glu	4.9	4.7	0.0	6.0	2.1	7.9	1.5	9.5
Gly	7.1	7.9	4.6	9.0	4.2	6.9	6.1	7.6
His	2.4	2.8	3.1	1.0	0.0	0.9	4.6	3.8
Ile	2.8	3.1	0.0	1.0	0.0	6.9	1.5	4.7
Leu	4.7	5.0	4.6	8.0	1.0	7.9	0.0	3.8
Lys	2.8	3.8	0.0	1.0	4.2	0.9	1.5	4.7
Met	0.7	0.3	0.0	1.0	1.0	0.0	1.5	0.9
Phe	2.7	3.5	3.1	1.0	1.0	3.9	1.5	2.8
Pro	10.5	8.5	23.4	7.0	16.8	6.9	13.8	7.6
Ser	8.8	4.7	20.3	7.0	26.3	4.9	6.1	4.7
Thr	11.1	5.7	20.3	4.0	26.3	6.9	32.3	4.7
Trp	1.2	1.2	1.5	2.0	0.0	1.9	0.0	1.9
Tyr	2.4	3.8	0.0	3.0	0.0	3.9	0.0	1.9
Val	6.1	6.6	6.2	7.0	5.2	3.9	4.6	7.6

and -3c peptides (Figures 4c and 4d). A putative signal sequence or polyadenylation signal was not present in HGM-1, indicating that this clone represents an internal fragment of the HGM cDNA. A tandemly repeated amino acid sub-sequence was not present either. Two consensus sequences for N-linked glycosylation were found at positions 19 and 304.

The amino acid composition of HGM-1 is typical of mucin glycoproteins, with large amounts of hydroxy amino acids and proline. As anticipated, cysteine also accounts for a considerable number of the amino acid residues (Table 2). Stretches of three subsequent hydroxy amino acids (e.g. TSS SST STS) were found 13 times in HGM-1. The hydroxy amino acids and proline appeared to be enriched in three clusters of 64, 65 and 95 residues. Each of these clusters contained four copies of sequence TXXP, which has been implicated as a major site for GalNAc addition [41–44]. These regions are likely to be heavily glycosylated. The amino acid sequence of HGM-1 could be subdivided as indicated in Figure 3 into seven domains differing considerably in amino acid composition (Table 2). Noticeably, cysteine-rich domains were separated by domains with extremely high threonine, serine and proline contents.

The cysteine-rich domain I (313 amino acid residues) had high sequence similarity to the D-domains of human vWF and MUC2, and exemplifies part of a putative D-domain in the HGM gene. The highest similarity was observed when this sequence was compared with the MUC2 D3-domain, the D-domain located just upstream of the tandem repeat region in the *MUC2* gene; 52% of the amino acid residues were identical (Figure 4a). The three other cysteine-rich domains (domains III, V and VII) were aligned using the method of Myers and Miller [45] (Figure 4b). Interestingly, several amino acid residues, especially two tryptophan and 10 cysteine residues, were exactly conserved in all three domains. This identity was also found when the sequences were aligned with similar domains in TBM and MUC2, but not vWF (Figure 4b).

DISCUSSION

We previously developed methods to obtain highly purified and intact HGM and generated specific antibodies [12–14]. We have now utilized this method to isolate HGM for the generation of unique nucleotide probes. In dot-blot experiments the isolated mucin reacted specifically and intensely with the antisera previously raised against HGM and RGM. A tryptic fragment of HGM was sequenced and appeared identical to a previously reported sequence of TBM [32]. Clone HGM-1 contained the nucleotide sequence encoding this peptide, thus providing evidence that this cDNA indeed represents part of the HGM gene. Additional evidence was supplied by Northern blotting studies indicating that the HGM cDNA probes hybridized to a very large transcript present in stomach tissue.

Identity of HGM cDNA

The 3′ region of clone HGM-1 clearly resembles part of the TBM cDNA (Figure 4b) [22,23]. Moreover, the 24 bp MUC5c repeat region was previously shown to be expressed both in the stomach and in the trachea [46]. This suggests that HGM-1, TBM cDNA and the MUC5c repeat region may be part of the same gene, which is expressed in the stomach, trachea and endocervix [22,23,46]. A consequence of the expression of this gene in the stomach and trachea would be that the HGM-t11/TBM:TR-3a fragment is present at least twice in the mucin, as the coding stretch of nucleotides is highly similar but not identical to the sequence of TBM (Figure 4c). Moreover, additional non-identical sequences coding for the same fragment were obtained by reverse transcriptase PCR [23] and by conventional cDNA cloning [22] (Figure 4c). Also, the TBM:TR-3b peptide was not exactly recovered as described previously (Figure 4d). Therefore, if HGM-1 indeed represents an upstream part of TBM cDNA, the gene product would contain repeated highly similar cysteine-rich

regions. Expression of the same gene in stomach and trachea would pose interesting questions concerning the transcriptional regulation of this gene. However, we cannot exclude the possibility that HGM-1 represents a novel mucin gene, which has high sequence similarity to TBM cDNA and is expressed exclusively in the stomach. To discriminate between the two possibilities, the availability of the full-length sequences of mucin genes isolated independently from trachea and stomach will be necessary. Until this has been clarified, we will refer to our clone as HGM cDNA.

Similarity of HGM-1 cDNA with cysteine-rich D-domains

Analysis of clone HGM-1 revealed some interesting features with regard to the structure of gastric mucin and to the architecture of mucin genes in general. First, we identified the N-terminal 264 residues as part of a putative D-domain, analogous to similar domains found in FIM-B1, BSM, PSM, TBM, MUC2, the rat homologue of MUC2 and vWF [19,23,26–28,31,47,48]. The two domains at the N-terminus of vWF are responsible for the conversion of vWF dimeric precursors into higher oligomers. These domains are proteolytically cleaved from the remainder of the molecule during biosynthesis [49–51]. A similar function in oligomerization has been suggested for the D-domains found in mucins, and the presence of a D-domain in HGM cDNA suggests that this function also applies to HGM. In fact, it was recently shown that MUC2 polypeptide precursors form intermolecular disulphide-bond-mediated oligomers during synthesis [52]. Furthermore, gastric mucin precursors in rat and man form oligomers. This process was proposed to occur in the ER of mucin-synthesizing cells, before addition of the first O-linked sugar chains. Since gastric mucin precursor oligomers include trimers and tetramers, the subunits can form intermolecular disulphide bonds at both ends of the molecule [13,14]. It was proposed that the final number of covalently linked mucin precursors is determined in the ER. In contrast, the ER form of the vWF precursors is exclusively monomeric or dimeric. Higher oligomers are formed in a later, slightly acidic, biosynthetic compartment [49–51]. Although HGM cDNA may contain at least one D-domain, these data indicate that the oligomerization of mucin precursors differs from the polymerization of vWF subunits.

Domain structure of HGM-1

In addition to the putative D-domain, three other similar cysteine-rich domains were identified in HGM-1. Again, the spatial arrangement of cysteine residues was strictly conserved in two similar domains in MUC2 and in one identified domain in TBM (Figure 4), suggesting a functional importance of these domains for gel-forming mucins. Similar structures in other mucin partial cDNAs or in vWF were not found. Whether these domains exhibit specific functions in HGM biosynthesis or whether they interact with other components of the mucus gel is as yet unknown.

Another interesting aspect of the deduced polypeptide sequence is the alternating arrangement of four cysteine-rich and three small TSP-rich clusters (Figure 5). At this time, there is no indication that the number of these alternating clusters is limited to the seven that are described here. The TSP-rich clusters contain no cysteines but very high amounts of serine + threonine, and are likely to be extensively O-glycosylated. Interestingly, they do not seem to have a tandem repeat organization. The O-linked glycans and the high number of proline residues presumably cause an extended conformation of these domains [3]. The cysteine-rich domains are most probably glycosylated to a much lesser extent and may form more globular structures due to disulphide bond formation. Therefore, in HGM, highly glycosylated domains may alternate with more globular modules. A similar organization was described in integumentary mucin C1 of *Xenopus laevis*, which contains cysteine-rich P-domains followed by consecutive semi-repetitive threonine-rich clusters [31]. However, no direct homology between the described cysteine-rich domains of HGM-1 and P-domains is apparent. An alternating cluster structure has also been proposed by Sheehan and Carlstedt for cervical mucins [53]. This was based on electron microscopic imaging of these molecules and tryptic fragments of them after binding of colloidal gold particles to supposed hydrophobic, non-glycosylated, domains of the mucins. The gold-labelled domains were located approx. 200 amino acid residues apart, which is comparable to the number of residues encompassed by one cysteine-rich domain and one TSP-rich region. Therefore, although differences may exist between cervical and gastric mucins, the observed oligosaccharide clusters may be reflected by the alternating domain structure of the mucin cDNA.

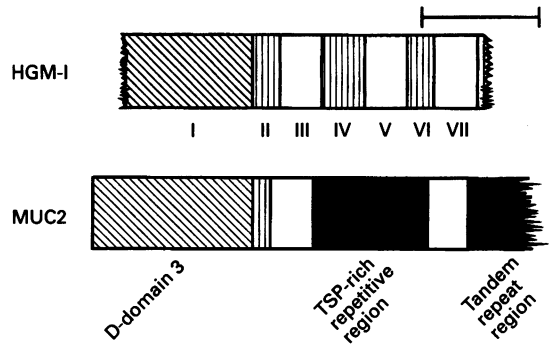

Figure 5 Representation of the HGM-1 domain structure

The seven identified domains of HGM-1 are depicted and compared with the organization of the most similar region in the deduced MUC2 gene product, immediately N-terminal to the tandem repeat region [19]. ▧, Putative D-domains; ▥, indicate TSP-rich regions; □, cysteine-rich domains with conserved position of cysteines. The semi-repetitive region and exact tandem repeats of MUC2 are in black. The bar indicates 250 amino acid residues.

The similarity of the HGM-1 sequence with that of the MUC2 cDNA (Figures 4 and 5) suggests that the gastric and intestinal mucin genes may have a common ancestor. The sequence of the third D-domain of MUC2 is more similar to the incomplete D-domain of HGM-1 than to the other MUC2 D-domains. This suggests that, during evolution, duplication of the D-domains in this putative ancestor gene preceded the formation of two independent genes. The described structural features may apply to more members of the mucin gene family. In conclusion, these data indicate that the HGM gene product contains at least one D-domain, probably involved in intermolecular disulphide bond formation, and in addition three other cysteine-rich domains with unknown function. The cysteine-rich domains alternated with short highly glycosylated modules.

We gratefully acknowledge Dr. H. Obertop and Dr. L. M. A. Akkermans (Department of Surgery, Academic Hospital, Utrecht) for their kind help in obtaining human tissue, and Dr. Alan Schwartz and Dr. Jonathan Gitlin for help with the RNA isolations and for the amino acid sequencing of HGM, which was performed at Monsanto, St. Louis, MO, U.S.A. We thank Tom van Rijn and René Scriwanek for preparation of the figures, and Dr. P. van der Sluijs for critical comments on the manuscript. The research described in this paper is supported by the Dutch Organisation for Scientific Research (NWO; grant 900-522-085) and NATO (grant 870316).

REFERENCES

1 Neutra, M. R. and Forstner, J. F. (1987) in Physiology of the Gastrointestinal Tract (Johnson L. R., ed.), pp. 975–1009, Raven Press, New York
2 Strous, G. J. and Dekker, J. (1992) Crit. Rev. Biochem. Mol. Biol. **27**, 57–92
3 Jentoft, N. (1990) Trends Biochem. Sci. **15**, 291–294
4 Allen A, (1981) in Physiology of the Gastrointestinal Tract (Johnson, L. R., ed.), pp. 617–639, Raven Press, New York
5 Forstner, J. F., Jabbal, I., Qureshi, R., Kells, D. I. C. and Forstner, G. G. (1979) Biochem. J. **181**, 725–732
6 Slayter, H. S., Lamblin, G., LeTreut, A., Galabert, C., Houdret, N., Degand, P. and Roussel, P. (1984) Eur. J. Biochem. **142**, 209–218
7 Harding, S. E., Rowe, A. J. and Creeth, J. M. (1983) Biochem. J. **209**, 893–896
8 Thornton, D. J., Davies, J. R., Kraayenbrink, M., Richardson, P. S., Sheehan, J. K. and Carlstedt, I. (1990) Biochem. J. **265**, 179–186
9 Dekker, J., van der Ende, A., Aelmans, P. H. and Strous, G. J. (1991) Biochem. J. **279**, 251–256
10 Dekker, J., Van Beurden-Lamers, W. M. O., Oprins, A. and Strous, G. J. (1989) Biochem. J. **260**, 717–723
11 Dekker, J., Aelmans, P. H. and Strous, G. J. (1991) Biochem. J. **277**, 423–427
12 Dekker, J., van Beurden-Lamers, W. M. O. and Strous G. J. (1989) J. Biol. Chem. **264**, 10431–10437
13 Dekker, J. and Strous, G. J. (1990) J. Biol. Chem. **265**, 18116–18122
14 Klomp, L. W. J., Van Rens, L. and Strous, G. J. (1994) Biochem. J. **304**, 693–698
15 Ligtenberg, M. J. L., Vos, H. L., Gennissen, A. M. C. and Hilkens, J. (1990) J. Biol. Chem. **265**, 5573–5578
16 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
17 Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M. and Kim, Y. S. (1991) J. Clin. Invest. **88**, 1005–1013
18 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
19 Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Sidiki, B. and Kim, Y. S. (1994) J. Biol. Chem. **269**, 2440–2446
20 Gum, J. R., Hicks, J. W., Swallow, D. M., Lagace, R. L., Byrd, J. C., Lamport, D. T. A., Siddiki, B. and Kim, Y. S. (1990) Biochem. Biophys. Res. Commun. **171**, 407–415
21 Porchet, N., Vancong, N., Dufosse, J., Audie, J. P., Guyonnet-Duperat, V., Gross, M. S., Deni, C., Degand, P., Bernheim, A. and Aubert, J. P. (1991) Biochem. Biophys. Res. Commun. **175**, 414–422
22 Aubert, J. P., Porchet, N., Crepin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D. and Degand, P. (1991) Am. J. Respir. Cell. Mol. Biol. **5**, 178–185
23 Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M. and Rose, M. C. (1994) J. Biol. Chem. **269**, 12932–12939
24 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
25 Tytgat, K. M. A. J., Buller, H. A., Opdam, F. J. M., Kim, Y. S., Einerhand, A. W. C. and Dekker, J. (1994) Gastroenterology **107**, 1352–1363
26 Xu, G., Huan, L.-J., Khatri, I. A., Wang, D., Bennick, A., Fahim, R. E. F., Forstner, G. G. and Forstner, J. F. (1992) J. Biol. Chem. **267**, 5401–5407
27 Eckhardt, A. E., Timpte, C. S., Abernethy, J. L., Zhao, Y. and Hill, R. L. (1991) J. Biol. Chem. **266**, 9678–9686
28 Bhargava, A. K., Woitach, J. T., Davidson, E. A. and Bhavanandan, V. P. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 6798–6802
29 Hoffman, W. (1988) J. Biol. Chem. **263**, 7686–7690
30 Probst, J. C., Hauser, F., Joba, W. and Hoffmann, W. (1992) J. Biol. Chem. **267**, 6310–6316
31 Hauser, F. and Hoffmann, W. (1992) J. Biol. Chem. **267**, 24620–24624
32 Rose, M. C., Kaufman, B. and Martin, B. M. (1989) J. Biol. Chem. **264**, 8193–8199
33 Rijnboutt, S., Aerts, H. M. F. G., Geuze, H. J., Tager, J. M. and Strous, G. J. (1991) J. Biol. Chem. **266**, 4862–4868
34 Carlstedt, I., Lingren, H., Sheehan, J. K., Ulmsten, U. and Wingerup, L. (1983) Biochem. J. **211**, 13–22
35 Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) Biochemistry **18**, 5294–5299
36 Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. U.S.A. **69**, 1408–1412
37 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
38 Del Sal, G., Manfioletti, G. and Schneider, C. (1989) Biotechniques **7**, 514–519
39 Maeda, M., Oshiman, K.-O., Tamura, S. and Futai, M. (1990) J. Biol. Chem. **265**, 9027–9032
40 Taggart, R. T., Cass, L. G., Mohandas, T. K., Derby, P., Barr, P. J., Pals, G. and Bell, G. I. (1989) J. Biol. Chem. **264**, 375–379
41 Wilson, I. B. H., Gavel, Y. and Von Heijne, G. (1991) Biochem. J. **275**, 529–534
42 O'Connell, B. C., Hagen, F. K. and Tabak, L. A. (1992) J. Biol. Chem. **267**, 25010–25018
43 Elhammer, A. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G. and Kezdy, F. J. (1993) J. Biol. Chem. **268**, 10029–10038
44 Nishimori, I., Johnson, N. R., Sanderson, S. D., Perini, F., Mountjoy, K., Cerny, R. L., Gross, M. L. and Hollingsworth, M. A. (1994) J. Biol. Chem. **269**, 16123–16130
45 Myers, E. W. and Miller, W. (1988) Comput. Appl. Biosci. **4**, 11–17
46 Audie, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B. and Aubert, J. P. (1993) J. Histochem. Cytochem. **41**, 1479–1485
47 Bonthron, D. T., Handin, R. I., Kaufman, R. J., Wasley, L. C., Orr, E. C., Mitsock, L. M., Ewenstein, B., Loscalzo, J., Ginsburg, D. and Orkin, S. H. (1986) Nature (London) **324**, 270–273
48 Shelton-Inloes, B. B., Titani, K. and Sadler, J. E. (1986) Biochemistry **25**, 3164–3171
49 Verweij, C. L., Hart, M. and Pannekoek, H. (1987) EMBO J. **6**, 2885–2890
50 Mayadas, T. N. and Wagner, D. D. (1991) Ann. N.Y. Acad. Sci **614**, 153–166
51 Mayadas, T. N. and Wagner, D. D. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 3531–3535
52 McCool, D. J., Forstner, J. F. and Forstner, G. G. (1994) Biochem. J. **302**, 111–118
53 Sheehan, J. K. and Carlstedt, I. (1990) Biochem. J. **265**, 169–178

Received 4 November 1994/9 February 1995; accepted 16 February 1995