Bovine gall-bladder mucin contains two distinct tandem repeating sequences: evidence for scavenger receptor cysteine-rich repeats

David P. NUNES, Andrew C. KEATES, Nezam H. AFDHAL and Gwynneth D. OFFNER*
Section of Gastroenterology, Department of Medicine, Boston University School of Medicine and Boston City Hospital, Boston, MA 02118, U.S.A.

Gall-bladder mucin is a densely glycosylated macromolecule which is the primary secretory product of the gall-bladder epithelium. It has been shown to bind cholesterol and other biliary lipids and to promote cholesterol crystal nucleation *in vitro*. In order to understand the molecular basis for mucin–lipid interactions, bovine gall-bladder mucin cDNAs were identified by expression cloning and were isolated and sequenced. The nucleotide sequences of these cDNAs revealed two distinct tandem repeating domains. One of these domains contained a 20-amino acid tandem repeating sequence enriched in threonine, serine and proline. This sequence was similar to, but not identical with, the short tandem repeating sequences identified previously in other mammalian mucins. The other domain contained a 127-amino acid tandem repeating sequence enriched in cysteine and glycine. This repeat displayed considerable sequence similarity to a family of receptor- and ligand-binding proteins containing scavenger receptor cysteine-rich repeats. By analogy with other proteins containing these cysteine-rich repeats, it is possible that, in gall-bladder mucin, this domain serves as a binding site for hydrophobic ligands such as bilirubin, cholesterol and other biliary lipids.

INTRODUCTION

Mucins are complex glycoproteins secreted by epithelial cells and found on the luminal surface of the respiratory, gastrointestinal and genitourinary tracts. These proteins form a barrier to physical, chemical and bacterial injury by lubricating mucosal surfaces and by preventing direct cellular contact with a large variety of noxious agents and organisms [1–3].

Owing to their large size and extensive glycosylation, direct determination of the primary structure of mucins has been difficult and most of the information available on the sequence of the polypeptide backbone has been deduced from the nucleotide sequences of cDNA clones. Analysis of the partial or complete sequences of several distinct human mucins, MUC1–MUC7 [4–17], as well as mucins from other species [18–27], has revealed several characteristic features common to almost all of these proteins. First, tandem repeating sequences which, in most cases, are rich in proline, serine and threonine, are thought to comprise the central region of the protein backbone. Secondly, cysteine-rich N-terminal [16] and C-terminal [8,16,17,19,21,22,24,26] domains have been shown to flank the heavily glycosylated central tandem repeat region. These cysteine-rich domains may be important for the formation of disulphide bonds between mucin monomers, a feature that is thought to be required for the formation of mucin gels.

Gall-bladder mucin is the primary secretory product of gall-bladder epithelial cells and the major component of the mucus gel layer adherent to the gall-bladder epithelium. In addition to its cytoprotective function, gall-bladder mucin plays a key role in the pathogenesis of cholesterol gallstone disease [28,29]. Cholesterol crystal nucleation, the initial event in stone formation, is accelerated by gall-bladder mucin *in vitro* [30], while *in vivo*, nucleation and growth of cholesterol crystals is believed to occur within the mucin gel layer [31]. Furthermore, in the prairie dog, mucin hypersecretion occurs before gallstone formation [32] and inhibition of mucin secretion with aspirin can prevent the occurrence of stones [33].

Earlier studies from this laboratory have shown that gall-bladder mucin, like other mucins, contains two distinct structural domains, a serine- and threonine-rich glycosylated domain and a protease-sensitive non-glycosylated domain that contains binding sites for cholesterol and other hydrophobic ligands [30]. The present work was undertaken to obtain structural information on these domains in order to understand further the nature of the interaction(s) between mucin and biliary lipids.

EXPERIMENTAL

Antibody preparation

Mucin was isolated from bovine gall-bladders as previously described and deglycosylated by HF treatment [34]. The deglycosylated mucin was used to immunize male New Zealand White rabbits and IgG was purified from the resulting antisera using Affi-gel Blue (Bio-Rad Laboratories, Rockville Center, NY, U.S.A.) and titred against the deglycosylated mucin [34].

RNA isolation

Bovine gall-bladder, liver, stomach, small intestine, large intestine, heart and lung were obtained from a local abattoir and frozen immediately in liquid nitrogen. RNA was isolated as described by Chomczynski and Sacchi [35]. For library construction, bovine gall-bladder mRNA was isolated by affinity chromatography on oligo(dT)-cellulose (Pharmacia, Piscataway, NJ, U.S.A.).

cDNA library preparation and screening

A random-primed bovine gall-bladder cDNA library in λZap II (Stratagene, La Jolla, CA, U.S.A.) was prepared according to the manufacturer's protocol except that random hexamers

Abbreviations used: 1 × SSC, 0.15 M NaCl/0.015 M sodium citrate; SRCR domain, scavenger receptor cysteine-rich domain.
* To whom correspondence should be addressed.

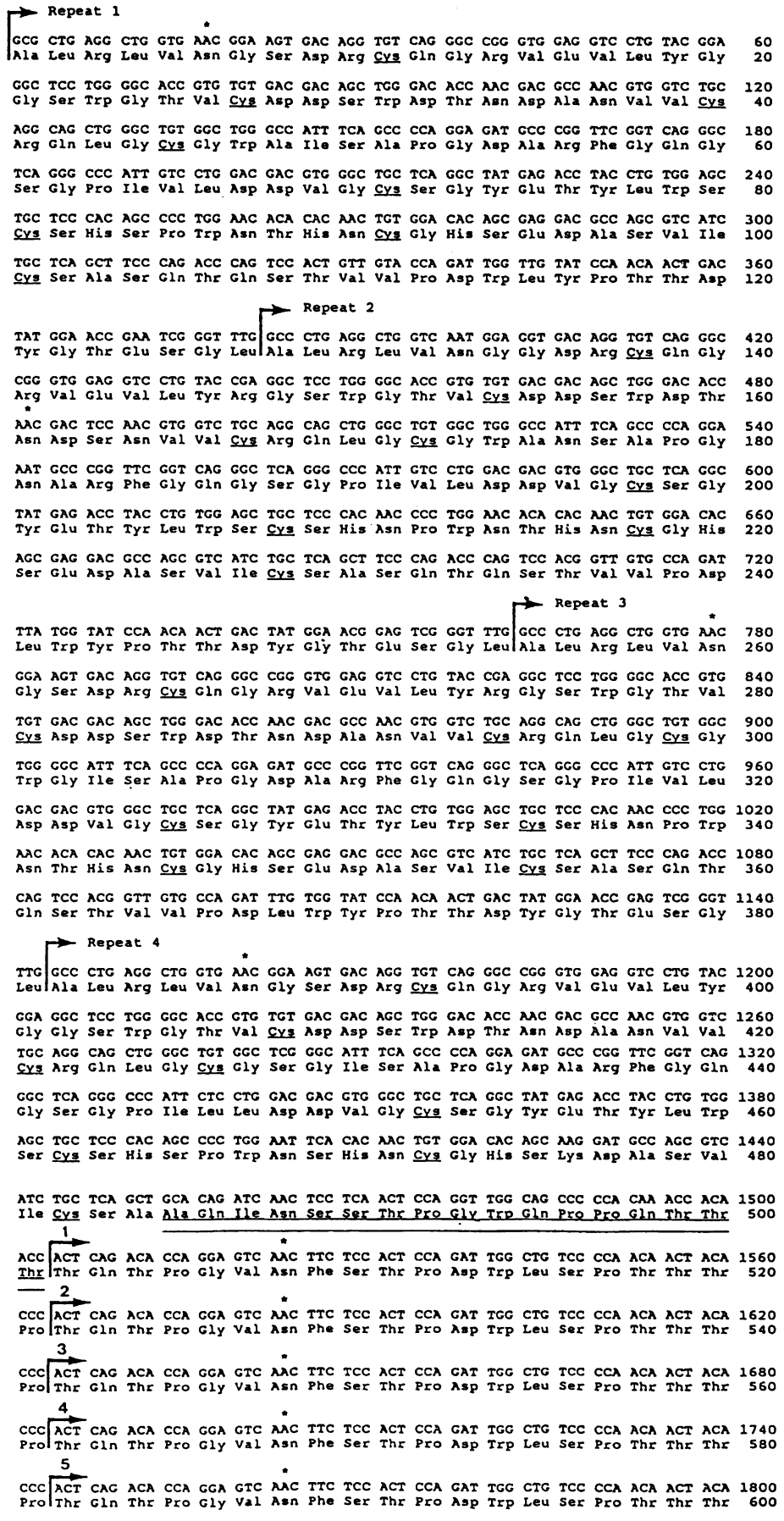

Figure 1 For legend see opposite.

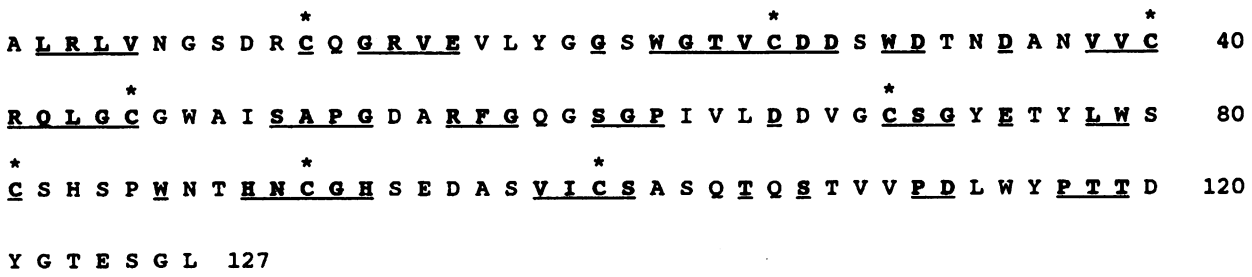

Figure 2 Consensus amino acid sequence of the 127-amino acid repeat in bovine gall-bladder mucin

The sequence of each of the repeats shown was compared with the sequences of all repeats found in the other clones isolated. Invariant residues are underlined and shown in bold type. The positions of the eight cysteine residues are marked with asterisks.

(Pharmacia) were used to prime first-strand cDNA synthesis. Approx. 200000 plaque-forming units were plated on *Escherichia coli* XL1-Blue at a density of 20000 plaque-forming units/150 mm Petri plate. After incubation at 42 °C for 3.5 h, plates were overlaid with nitrocellulose filters soaked in 10 mM isopropyl thiogalactopyranoside and incubated at 37 °C for a further 3 h. After being blocked, filters were incubated with a 1:600 dilution of the anti-(deglycosylated mucin) immune serum which had been treated with an *E. coli* lysate to remove antibodies cross-reacting with *E. coli* proteins. Filters were then incubated with alkaline phosphatase-conjugated goat anti-(rabbit IgG) (Promega Corp., Madison, WI, U.S.A.; 1:7500 dilution) and colour developed with 5-bromo-4-chloro-3-indolyl phosphate/Nitro Blue Tetrazolium. Positive clones were replated and rescreened until plaque purified.

Northern- and Southern-blot hybridization

RNA from bovine tissues (15 μg) was electrophoresed on 1% agarose denaturing gels and transferred to Hybond N^+ membranes (Amersham, Arlington Heights, IL, U.S.A.). Bovine genomic DNA (10 μg; Clontech, Palo Alto, CA, U.S.A.) was digested with a series of restriction enzymes and the digests were electrophoresed and blotted on to Hybond N^+ membranes. Northern and Southern blots were hybridized with random-primer-labelled [36] probes at 42 °C in a solution containing 25 mM potassium phosphate buffer, pH 7.4, 5 × SSC, 5 × Denhardt's, 100 μg/ml denatured salmon sperm DNA, 1% SDS, 50% formamide and 10% dextran sulphate (where 1 × SSC is 0.15 M NaCl/0.015 M sodium citrate and 1 × Denhardt's is 0.02% Ficoll 40/0.02% polyvinylpyrrolidone/0.02% BSA). Final washes were performed in 0.2 × SSC at 42 °C.

DNA sequencing

After purification of positive plaques, insert-containing pBluescript phagemids were excised from phage clones as described by the manufacturer (Stratagene). DNA was isolated from 25 of these clones and sequenced from both ends using the dideoxy method [37] with Sequenase v. 2.0 (United States Biochemical Corp., Cleveland, OH, U.S.A.). The complete sequences of several clones were determined from unidirectional deletions prepared using a commercially available exonuclease III system (Erase-a-base; Promega). Analysis of nucleotide and deduced protein sequences was performed using Intelligenetics Suite (Intelligenetics, Palo Alto, CA, U.S.A.) software.

RESULTS

Isolation and nucleotide sequences of gall-bladder mucin cDNAs

Fifty immunopositive plaques were identified by screening approx. 200000 plaque-forming units from the bovine gall-bladder cDNA library with anti-(deglycosylated mucin) immune serum, and 25 were plaque-purified and characterized. All 25 clones contained a 381 bp tandem repeating sequence which was present in two to four copies per clone and several also contained multiple copies of a 60 bp repeating sequence located 3′ to the 381 bp repeating sequence. The nucleotide and deduced amino acid sequence of one of the gall-bladder mucin cDNA clones, pGBM7-1, is shown in Figure 1.

pGBM7-1 contains three complete and one partial 381 bp tandem repeating sequences which each contain an open reading frame coding for 127 amino acids. In contrast with the repeats in other previously described mucins, this sequence contains relatively few serine and threonine residues (19%) and therefore has few potential O-glycosylation sites. However, each of the 127 amino acid repeats contains one potential N-glycosylation site (marked with asterisks in Figure 1). Interestingly, each of the repeats contains eight cysteine residues, four of which are equally spaced 10 amino acids apart. Comparison of the sequences obtained from all of the clones studied reveals that the pattern of the eight cysteine residues is invariant. As shown in the consensus sequence (Figure 2), 102 of the 127 amino acid residues in the tandem repeat are identical in all clones studied. Substitutions at the remaining 25 positions each occurred in more than one clone. This suggests that clones containing these conserved, yet distinct, tandem repeats are derived from different regions of the gall-bladder mucin mRNA.

At the 3′ terminus of the 381 bp tandem repeating sequence in pGBM7-1 is a 51 bp linker region coding for 17 amino acids, 11

Figure 1 Nucleotide and deduced amino acid sequence of bovine gall-bladder mucin cDNA clone pGBM7-1

The 127-amino acid repeats are indicated as repeats 1–4, the 17-amino acid linker region is double underlined and the 20-amino acid repeats are labelled 1–5. Cysteine residues are underlined and potential N-linked glycosylation sites are marked with an asterisk.

```
AGT GGG GCC CAA GTC AGT CCA AAG TTC ACG AGG TGG CCA GAA GGA GAC CCT GGC CTC CAG   60
Ser Gly Ala Gln Val Ser Pro Lys Phe Thr Arg Trp Pro Glu Gly Asp Pro Gly Leu Gln   20

AGC CTA TCA CCA GAC CCA GAG AGA AGC TAC AGG TCT TGG GCA AGC TCT AGC CAT CCA GGA  120
Ser Leu Ser Pro Asp Pro Glu Arg Ser Tyr Arg Ser Trp Ala Ser Ser Ser His Pro Gly   40

GCC CTT GGC AAA CTG CCT TCC TGT CCC AAG ACA CAC AGA TGT GCT TGG AGG AGC CAA GAA  180
Ala Leu Gly Lys Leu Leu Ser Cys Pro Lys Thr His Arg Cys Ala Trp Arg Ser Gln Glu   60

GAA AGC AGG AAA GTG TCC CTG GGG GAT AGA GGG CAG CCT AAA TAC ACC CTA ACT ACA CCC  240
Glu Ser Arg Lys Val Ser Leu Gly Asp Arg Gly Gln Pro Lys Tyr Thr Leu Thr Thr Pro   80
                              Region I
 → 1                *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC  300
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro  100
 → 2                *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC  360
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro  120
 → 3                *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC  420
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro  140
 → 4                *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC  420
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro  160
 → 5                *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC  540
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro  180
                              Region II
                    *
ACT CAG ACA CCA GGA GTC AAC TTC TCC ACT CCA GGC TCC TTT TCA AGT TGT GGT GGC TTC  600
Thr Gln Thr Pro Gly Val Asn Phe Ser Thr Pro Gly Ser Phe Ser Ser Cys Gly Gly Phe  200

TTA TTC AGT GGC AGT GGG AAC TTT TGT AGC CCA TCC TAC CCA GGA TAC TAC CCC AAC AAC  660
Leu Phe Ser Gly Ser Gly Asn Phe Cys Ser Pro Ser Tyr Pro Gly Tyr Tyr Pro Asn Asn  220

GCC GAC TGT GTC TGG GAA ATA CAA GTG AAC CCC GGC TAC CTC GAT AAC CTG GGC TTC GAC  720
Ala Asp Cys Val Trp Glu Ile Gln Val Asn Pro Gly Tyr Leu Asp Asn Leu Gly Phe Asp  240

AGT CTG CAG TTG GAG ACA CAC AGT AGC TGC AGT TAT GAC TAT GTT GAA ATC CTT AAT GGA  780
Ser Leu Gln Leu Glu Thr His Ser Ser Cys Ser Tyr Asp Tyr Val Glu Ile Leu Asn Gly  260
                *
CCG CTG AGT AGC AAT GCC TCA GCG AGG AGA ATC TGT CTG TAC ACC AGG GAA ATA TTC ACT  840
Pro Leu Ser Ser Asn Ala Ser Ala Arg Arg Ile Cys Leu Tyr Thr Arg Glu Ile Phe Thr  280

TCT TAT TCC AAC CGA TTG ACT GTT CGA TTT CGG AGT GAC GGC AGT GTC CAA AAA ACT GGT  900
Ser Tyr Ser Asn Arg Phe Thr Val Arg Phe Arg Ser Asp Gly Ser Val Gln Lys Thr Gly  300
                                                *                           *
TTT TCT GCT TGG TAT AAC TCC TTT CCA AGA AAT GTC AGC TTG AGA TTG GTG AAC TGG AAC  960
Phe Ser Ala Trp Tyr Asn Ser Phe Pro Arg Asn Val Ser Leu Arg Leu Val Asn Trp Asn  320
                              Region III

TCC TCC CAT CCC ACA TGT GCT GGG CGT GTG GAA ATC TAC CAT GGT GGC CAG TGG GGA ACA 1020
Ser Ser His Pro Thr Cys Ala Gly Arg Val Glu Ile Tyr His Gly Gly Gln Trp Gly Thr  340

GTG TGC GAT GAC AAC TGG GAC GTT CAA GAT GCC CAG GTG GTG TGC AGA CAG CTG GGC TGT 1080
Val Cys Asp Asp Asn Trp Asp Val Gln Asp Ala Gln Val Val Cys Arg Gln Leu Gly Cys  360

GGA TAT GCA GTC TCA GCC CCT GGA AAT GCC TAC TTT GGC TCT GGC TCT GGT CCC ATC ACC 1140
Gly Tyr Ala Val Ser Ala Pro Gly Asn Ala Tyr Phe Gly Ser Gly Ser Gly Pro Ile Thr  380

TTG GAT GAC GTG GTG TGC TCA GGG GCG GAG TCC AAT CTC TGG CAG TGC CGG AAC CGA GGA 1200
Leu Asp Asp Val Val Cys Ser Gly Ala Glu Ser Asn Leu Trp Gln Cys Arg Asn Arg Gly  400

TGG TTC TAC CAC AAT TGT GGC CAC CAT GAA GAT GCT GGA GTC ATT TGC TCA GAT ATA CCG 1260
Trp Phe Tyr His Asn Cys Gly His His Glu Asp Ala Gly Val Ile Cys Ser Asp Ile Pro  420

    *
ACC AAC TCC TCC ACT CCA GAT TGG CTG TCC CCA ACA ACT ACA CCC ACT CAG AAT CCT GAT 1320
Thr Asn Ser Ser Thr Pro Asp Trp Leu Ser Pro Thr Thr Thr Pro Thr Gln Asn Pro Asp  440

                              Region IV
    *                                                                   *
TAC AAC GTC ACC GGC CCC AGC CAC TGC GGA GGC TTC CTG ACC CAG TTT TCA GGG AAC TTT 1380
Tyr Asn Val Thr Gly Pro Ser His Cys Gly Gly Phe Leu Thr Gln Phe Ser Gly Asn Phe  460

TCC AGC CCA TTC TAC CCT AGG AAC TAT CCG AAC AAC GCC AAG TGT GTG TGG GAC ATT GAA 1440
Ser Ser Pro Phe Tyr Pro Arg Asn Tyr Pro Asn Asn Ala Lys Cys Val Trp Asp Ile Glu  480
        *
GTT CAA AAC CAC AGT TCC CCG CTG CTC GCC CGG GTT TGT GAC GGG TCA AGG GGC TCC TTC 1500
Val Gln Asn His Ser Ser Pro Leu Leu Ala Arg Val Cys Asp Gly Ser Arg Gly Ser Phe  500

ACC TCA TCG TCC                                                                  1512
Thr Ser Ser Ser               Region V
504
```

Figure 3 Nucleotide and deduced amino acid sequence of bovine gall-bladder mucin clone pGBM31-1

The sequence has been divided into regions I–V where region II contains the 20-amino acid tandem repeats (labelled 1–5), regions III and V contain a complement C1r-like sequence and region IV contains a 131-amino acid cysteine-rich sequence (underlined). Potential N-glycosylation sites are marked with an asterisk.

of which are serine, threonine or proline. This sequence connects to a region containing five copies of a 60 bp tandem repeating sequence. Some 40 % of the amino acids in the deduced sequence of this tandem repeat are either threonine or serine, indicating that this region contains multiple potential O-glycosylation sites. In addition, one potential N-glycosylation site is present in each repeat unit (Figure 1). The clones in which this sequence occurred contained between four and eight identical tandem repeat units. This 20-amino acid tandem repeating sequence is similar to, but not identical with, the repeating sequences previously described in other gastrointestinal mucins.

In all of the clones containing both repeat units, the 381 bp repeats were located 5′ to the 60 bp repeats and therefore an attempt was made to identify sequences occurring 3′ to the latter repeats. Differential screening of the cDNA library was carried out using hybridization probes each coding for only one of the repeat units. One plaque filter was hybridized with a 381 bp *Pst*I fragment of pGBM7-1 containing three individual 381 bp repeat units (base 122–502, 503–883 and 884–1264; Figure 1). The other plaque filter was hybridized with a 305 bp *Bbv*I–*Eco*RI fragment containing five contiguous 60 bp repeats (base 1495–1800; Figure 1). Ten clones that hybridized exclusively to the 60 bp repeat probe were selected for further study.

The sequence of one of these clones, pGBM31-1, shown in Figure 3, reveals features that were common to all ten clones. Each contained five complete 60 bp repeat units with nucleotide sequences identical with those found in pGBM7-1 (region II, Figure 3). In addition, each contained one copy of a 393 bp sequence coding for a 131-amino acid sequence which contained eight cysteine residues (region IV, Figure 3). These cysteines could be perfectly aligned with the eight cysteine residues present in the 127-amino acid consensus repeat. Overall, the 131-amino acid sequence in pGBM31-1 displayed 50 % sequence identity with the 127-amino acid consensus sequence in pGBM7-1 (see Figure 4).

The sequence of pGBM31-1 also reveals several additional features which are unique to this clone. First, it contains a 240 bp sequence located 5′ to the 60 bp repeat units coding for a

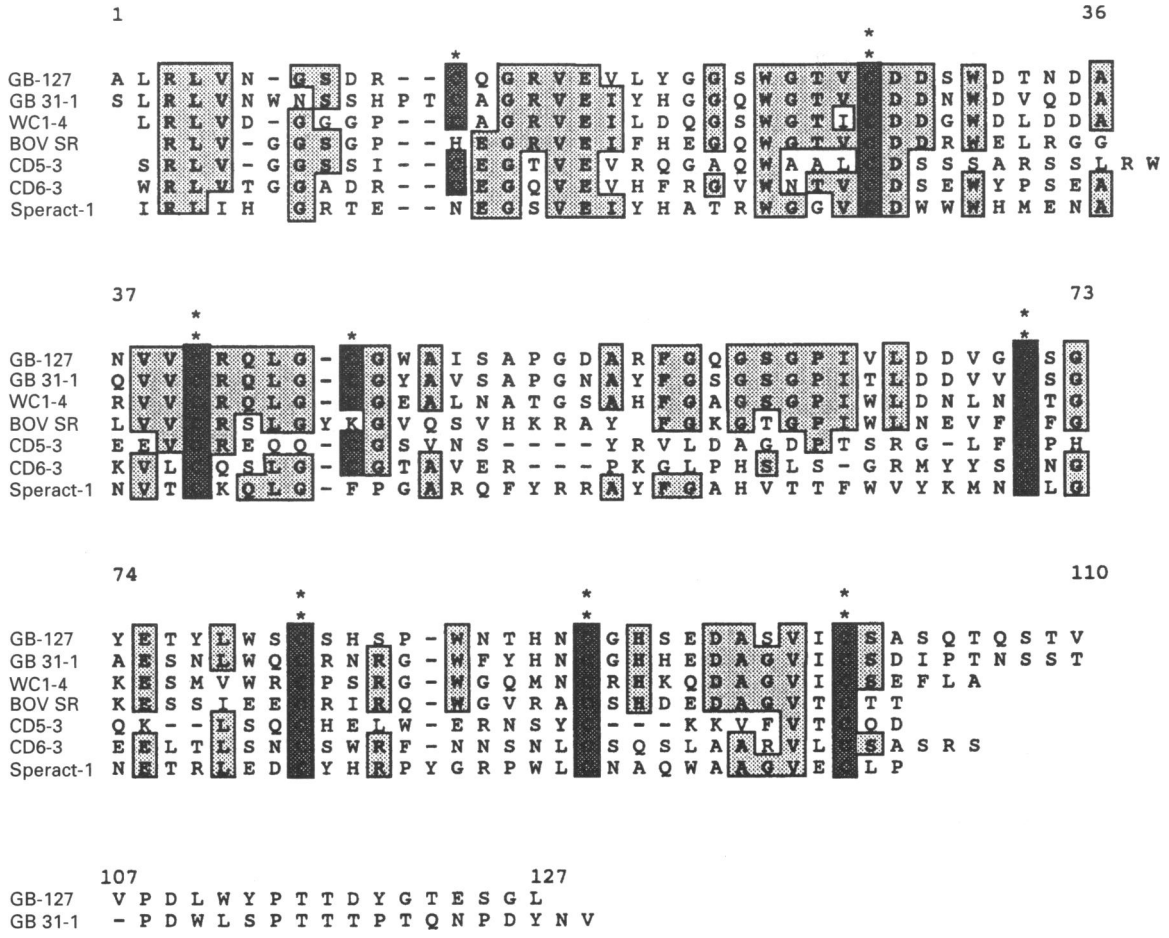

Figure 4 Comparison of the bovine gall-bladder mucin cysteine-rich repeat sequences with the sequences of five other proteins containing scavenger receptor cysteine-rich (SRCR) domains

Short gaps were introduced to maximize homology between the proteins. Amino acid residues identical in at least four of the seven sequences are enclosed in light grey boxes. Cysteine residues are enclosed in dark grey boxes. Abbreviations: GB-127, gall-bladder mucin 127-amino acid repeat consensus sequence; GB31-1, the 131-amino acid sequence in bovine gall-bladder mucin clone pGBM31-1; WC1-4, the fourth SRCR domain in the bovine $\gamma\delta$ T-lymphocyte surface WC1 antigen [38]; BOV SR, bovine macrophage scavenger receptor [39]; CD5-3, the third SRCR domain in human CD5 [40]; CD6-3, the third SRCR domain in human CD6 [41]; speract-1, the first SRCR domain in the sea-urchin speract receptor [42].

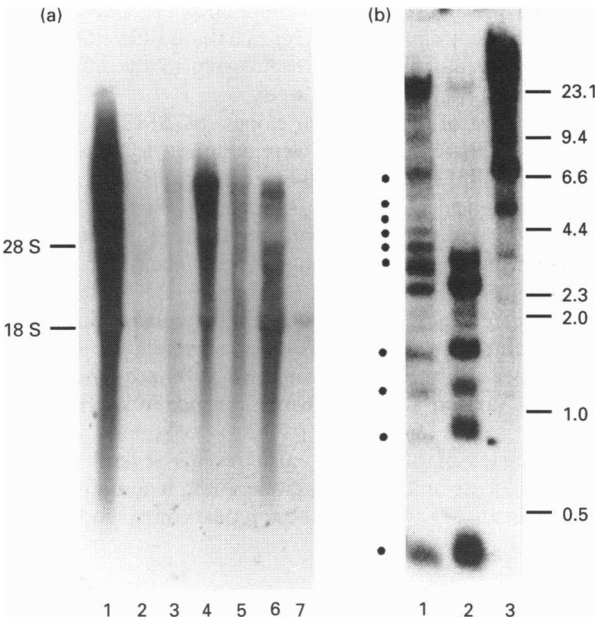

Figure 5 (a) Northern-blot analysis of RNAs from a variety of bovine tissues probed with a bovine gall-bladder mucin cDNA and (b) Southern-blot analysis of restriction-endonuclease-digested bovine genomic DNA hybridized with a 381 bp *Pst*I fragment coding for the 127-amino acid repeat

(**a**) RNAs were electrophoresed, blotted on to nylon membrane and hybridized with random-primer-labelled insert DNA from pGBM7-1 as described in the text. Lanes 1–7, RNA from gall bladder, stomach, small intestine, large intestine, lung, liver and heart respectively. (**b**) All lanes contain 10 μg of bovine genomic DNA cut with 10 units of restriction enzyme. Lane 1, partial digest (5 min) with *Pst*I; lane 2, complete (16 h) digest with *Pst*I; lane 3, complete digest with *Cfo*I. Size markers (kb) are indicated and bands that could be multiples of the 381 bp repeat are indicated with dots.

sequence in which serine, threonine and proline comprise 33 % of the amino acids (region I, Figure 3). This sequence could not be aligned with either of the repeating sequences described above and may represent a second type of glycosylated domain. In addition, pGBM31-1 contains a 385 bp sequence located between the 60 bp repeats and the 393 bp cysteine-rich sequence (region III, Figure 3). The deduced sequence of this region contains five cysteine residues and a search of the PIR database revealed a striking similarity to repeating domains found in complement component C1r and a number of other proteins [43–45]. Four of the five cysteine residues are conserved in these proteins. At the 3′ end of the 393 bp sequence is a 183 bp sequence (region V, Figure 3) which appears to code for another complement C1r-like sequence

Comparison of the 127-amino acid consensus repeat (Figure 2) and the 131-amino acid sequence from pGBM31-1 (Figure 3) with sequences in the PIR database revealed that the pattern of cysteine residues in these sequences is similar to that found in a number of other proteins with receptor- or ligand-binding functions (Figure 4). Those proteins with the greatest degree of sequence similarity include the WC1 protein expressed on bovine $\gamma\delta$ T-lymphocytes [38], bovine macrophage scavenger receptor [39], human T-lymphocyte glycoproteins CD5 [40,41] and CD6 [41] and sea-urchin speract receptor [42]. The position and spacing of six of the eight cysteine residues present in the gall-bladder mucin consensus sequence are conserved in each of these proteins (Figure 4). This pattern of conserved cysteine residues has been referred to as the scavenger receptor cysteine-rich (SRCR) domain [46]. In addition to the six conserved cysteine residues, eight other amino acids are also invariant among these sequences, ten amino acids are identical in six of the seven sequences and 13 amino acids are identical in five of the seven sequences.

Analysis of the secondary structure of the 127-amino acid gall-bladder mucin consensus repeat by the method of Chou and Fasman [47] predicted that this region was comprised exclusively of β-sheet and β-turn structures, with no helical structure (results not shown). This indicates that regions of the mucin molecule containing the 127-amino acid repeats probably assume an extended non-globular conformation. The hydropathicity plot of the 127-amino acid tandem repeating sequence revealed short, alternating hydrophilic and hydrophobic segments where the length of each segment ranges from five to 16 amino acid residues [48] (results not shown). Four of the eight cysteine residues are located in hydrophobic segments.

Tissue distribution of gall-bladder mucin mRNA

When the cDNA insert from pGBM7-1 was used to probe a Northern blot containing total RNA isolated from bovine gall bladder, stomach, small intestine, large intestine, lung, liver and heart, the strongest hybridization signal was detected in lanes containing gall-bladder RNA (Figure 5a). The diffuse pattern of hybridization extending from greater than 9 kb to approx. 1 kb is characteristic of that seen with other mucin mRNAs. Hybridization was also detected to a lesser extent in lanes containing RNA from large intestine, lung and liver, the latter probably representing mucin mRNA expression by biliary epithelial cells. No significant hybridization was detected in lanes containing RNA from stomach, heart or small intestine.

Southern-blot analysis of bovine genomic DNA

In order to obtain further information on the arrangement of the repeat units in this bovine gall-bladder mucin gene, Southern blots of digested bovine genomic DNA were hybridized with a probe containing only the 381 bp repeat (Figure 5b). When DNA was digested partially (lane 1) or completely (lane 2) with *Pst*I, a restriction enzyme that should cleave each of the 381 bp repeats once, a complex hybridization pattern was seen. The complete digest (lane 2) contains four major hybridizing bands at 380, 790, 1150 and 1500 bp. The 380 bp band should contain single repeat units and the larger bands may contain fragments with two, three and four repeat units respectively. Such fragments could be generated if the *Pst*I site was missing from some of the sequences containing the 381 bp repeat. It is noteworthy that the sequence of the *Pst*I site (CTGCAG) includes the last base of a valine codon (GTC), a codon for cysteine (TGC) and the first two bases of an arginine (or serine) codon (AGX). The sequence Val-Cys-Arg- is invariant in all clones examined (Figure 2); however, a change in the last base of the valine codon would eliminate the *Pst*I site, and give rise to the pattern observed. Two additional bands, at 2400 and 2900 bp, could represent fragments containing portions of the 381 bp repeat unit and its flanking sequences. The partial digest (lane 1) contained faint bands at 380, 790, 1150 and 1500 bp, identical in size with those in the complete digest. A ladder-like pattern of darker bands at approx. 2700, 3300, 3800, 4500, 4900 and 6300 bp was also seen. These bands could correspond to fragments containing partially cleaved multiples of the 381 bp repeat sequence with 7, 9, 10, 12, 13 and 16 individual repeat units respectively, or some may contain repeat units with flanking sequences. Bovine genomic DNA was also cleaved with *Cfo*I, a restriction enzyme with a four-base-recognition sequence for which there are no cleavage sites in the 381 bp repeat. *Cfo*I

digestion resulted in four major hybridizing bands at 3000, 4900, 6700 and 9120 bp (Figure 5b, lane 3). This hybridization pattern suggests that the repeat units are not tandemly arranged in the bovine gall-bladder mucin gene, but occur in separate domains, separated by sequences containing *Cfo*I sites.

Figure 6 Schematic representation of the possible organization of repeating units in bovine gall-bladder mucin

Black boxes correspond to domains containing the 20-amino acid repeats, grey boxes correspond to individual 127-amino acid repeats (or 131-amino acid sequences), and white boxes correspond to other sequences such as region I (Figure 3) or complement C1r-like domains.

DISCUSSION

Mucins secreted by epithelial cells lining the gastrointestinal and respiratory tracts function primarily in the protection of underlying tissue from mechanical, chemical or bacterial injury. However, in addition to its cytoprotective function, gall-bladder mucin has also been shown to be involved in the pathogenesis of cholesterol gallstone disease. Gall-bladder mucin binds hydrophobic ligands such as cholesterol and bilirubin and accelerates the nucleation of cholesterol crystals *in vitro* [30]. In the present paper, we present the first nucleotide sequences of mucin clones isolated from a gall-bladder cDNA library from any mammalian species and demonstrate a unique arrangement of structural units in this molecule.

The deduced amino acid sequence of a 1.8 kb clone, pGBM7-1, contained two different tandem repeating sequences, a 127-amino acid cysteine-rich repeat and a 20-amino acid repeat rich in serine, threonine and proline. Neither of the tandem repeating sequences has been identified in any mucin described previously. Further structural information was provided by the deduced amino acid sequence of a 1.5 kb clone, pGBM31-1, which contained both the 20 amino acid repeat and a 131-amino acid sequence which was similar to the 127-amino acid repeat found in the first clone. pGBM31-1 also had two non-contiguous copies of a cysteine-containing complement C1r-like repeat sequence. Thus bovine gall-bladder mucin contains at least three different repeating sequences.

In previous work, we have reported the amino acid sequences of four tryptic peptides derived from deglycosylated bovine gall-bladder mucin [34]. Although none of these peptides is contained within the 20-amino acid mucin-like repeat presented here, there are obvious similarities, both in amino acid composition (in threonine, proline and serine) and amino acid sequence. The sequence, Thr-Thr-Thr-Pro-Thr-Xaa-, where Xaa is either Val or Ser, occurs in three of the four peptides. The sequence, Thr-Thr-Thr-Pro-Thr-Gln-, occurs once in each of the tandem repeats (Figure 1). It is possible that the peptide sequences are derived from a region of the gall-bladder mucin molecule not yet identified in cloning studies. Such a region could contain degenerate tandem repeats, similar to those found in MUC2 [8,9].

In other mucins described to date, tandem repeating sequences enriched in serine, threonine and proline appear to occur in an uninterrupted array in the central portion of the molecule (reviewed in refs. [49,50]). Several lines of evidence suggest that bovine gall-bladder mucin has a unique structural organization, with the different repeating sequences arranged as a mosaic. First, in pGBM7-1 and several other clones, sequences coding for the 20-amino acid repeats occur 3′ to sequences coding for the 127-amino acid repeats. However, in pGBM31-1, the arrangement of the repeating units is reversed, with sequences coding for the 20-amino acid repeat located 5′ to sequences coding for the 131-amino acid sequence. Thus domains containing the 20-amino acid repeat appear to be interspersed with domains containing the 127-amino acid repeat or the 131-amino acid sequence. Secondly, Southern blots of bovine genomic DNA cleaved with *Cfo*I, a restriction enzyme for which there are no cleavage sites in the 381 bp repeat, revealed a complex pattern of bands when hybridized with the 381 bp repeat probe. Four major hybridizing bands, ranging in size from approx. 3.0 to 9.0 kb, were detected. This demonstrates that the 381 bp repeat units occur in distinct domains separated by other sequences. If the repeats were clustered in a single domain, cleavage with *Cfo*I, an enzyme with no cleavage sites within the repeat, would have produced a single large hybridizing band, as has been found for human MUC2 and MUC6 [9,13]. A schematic representation of the possible organization of these sequences in bovine gall-bladder mucin is shown in Figure 6.

The size of the individual repeat units can also be estimated from the Southern blot shown in Figure 5(b). Partial cleavage with *Pst*I, which should cleave once within every repeat, demonstrates that the largest repeat units contain 12–16 tandemly repeated copies of the 381 bp repeat. The smallest hybridizing band generated by cleavage with *Cfo*I was approx. 3.0 kb, suggesting that the smallest repeat units contain four to eight tandemly repeated copies of the 381 bp repeat.

Recent work on the characterization of human mucins has demonstrated that a given mucin-producing tissue expresses genes for more than one type of mucin (reviewed in refs. [49,50]). In the present work, all 25 of the bovine gall-bladder mucin cDNA clones isolated initially contained sequences coding for the 127-amino acid repeat, and several contained sequences coding for the 20-amino acid repeat as well. The failure to detect clones containing other mucin-like sequences could suggest that pGBM7-1 is representative of the major secretory mucin in the bovine gall bladder as the antibody used to screen the library should have been directed against all types of mucin present in bovine gall-bladder mucosal scrapings. However, differences in antigenicity may account for the preferential reactivity of this antibody to clones containing the 127- and/or 20-amino acid repeats.

The presence of repeating cysteine-rich sequences in bovine gall-bladder mucin is not unexpected as many previously described mucins contain cysteine-rich domains. The human intestinal mucin MUC2 has cysteine-rich regions as both the N- and C-termini [8,16]. The N-terminal region displays considerable sequence similarity with the D domains in human prepro-von Willebrand factor [51], and the C-terminal region can be aligned with the C-termini of von Willebrand factor [51], MUC5 [17], rat intestinal mucin-like protein [19], bovine and porcine submaxillary mucins [21,22] and frog integumentary mucin B1 [26]. It is likely that these cysteine-rich regions are involved in intermolecular disulphide bonding between mucin monomers. That disulphide linked oligomers are important in mucin gel formation has been suggested by the fact that treatment with reducing agents leads to disaggregation and gel depolymerization [52,53].

The cysteine-rich repeats in bovine gall-bladder mucin, in which the spacing of cysteine residues is nearly identical with that in the SRCR motif, have not been identified in any other mucin characterized to date. Other proteins that contain this domain are thought to be ligand-binding molecules. It has been suggested that the WC.1 protein on bovine $\gamma\delta$ T-lymphocytes may interact

with extracellular ligands, possibly antigen-presenting molecules, through the SRCR domains [38]. CD5, a T-lymphocyte surface glycoprotein that plays a role in T-cell activation, binds the B-cell surface glycoprotein CD72 [54], and the sea-urchin speract receptor serves as the receptor for the egg peptide speract [42]. The presence of SRCR domains in bovine gall-bladder mucin suggests that these regions may contain the hydrophobic binding sites identified previously.

The overall structural model of bovine gall-bladder mucin derived from the nucleotide sequence data presented here is in excellent agreement with that predicted from earlier biochemical characterization of this molecule. Bovine gall-bladder mucin was shown to contain two distinct domains, glycosylated regions rich in threonine and proline, and non-glycosylated regions rich in serine, glutamic acid/glutamine and glycine [55]. The latter domain contains numerous sites which bind cholesterol, phosphatidylcholine and hydrophobic fluors [30]. Binding of these ligands is abolished by protease digestion [30], but treatment with reducing agents, causing depolymerization into mucin monomers, appears to increase the number of available hydrophobic binding sites [56]. In this work, we show that bovine gall-bladder mucin is comprised of alternating glycosylated 20-amino acid repeats and non-glycosylated sequences such as the 127-amino acid SRCR repeats. The SRCR domains are non- or poorly glycosylated and contain approx. 31% serine, glutamic acid/glutamine and glycine. These domains would be expected to be protease-sensitive and, by analogy with other SRCR-containing proteins, might be expected to bind one or more ligands, possibly cholesterol or other biliary lipids. Furthermore the cysteine residues in the SRCR repeats are potentially capable of forming disulphide bonds between mucin monomers leading to polymerization and gel formation.

This work was supported by grant DK44619 from the National Institutes of Health.

REFERENCES

1 Strous, G. J. and Dekker, J. (1992) Crit. Rev. Biochem. Mol. Biol. **27**, 57–92
2 Neutra, M. and Forstner, J. F. (1987) in Physiology of the Gastrointestinal Tract, 2nd edn. (Johnson, L. F., ed.), pp. 975–1009, Raven Press, New York
3 Allen, A. and Pearson, J. P. (1993) Eur. J. Gastro. Hepatol. **5**, 193–199
4 Siddiqui, J., Abe, M., Hayes, D., Shani, E., Yunis, E. and Kufe, D. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2320–2323
5 Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J. et al. (1990) J. Biol. Chem. **265**, 15286–15293
6 Wreschner, D. H., Hareuveni, M., Tsarfaty, H. et al. (1990) Eur. J. Biochem. **189**, 463–473
7 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
8 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
9 Toribara, N. W., Gum, J. R., Culhane, P. J. et al. (1991) J. Clin. Invest. **88**, 1005–1013
10 Gum, J. R., Hicks, J. W., Swallow, D. M. et al. (1990) Biochem. Biophys. Res. Commun. **171**, 407–415
11 Porchet, N., Van Cong, N., Dufosse, J. et al. (1991) Biochem. Biophys. Res. Commun. **175**, 414–422
12 Aubert, J.-P., Porchet, N., Crepin, M. et al. (1991) Am. J. Respir. Cell Mol. Biol. **5**, 178–185
13 Toribara, N. W., Roberton, A. M., Ho, S. B. et al. (1993) J. Biol. Chem. **268**, 5879–5885
14 Dufosse, J., Porchet, N., Audie, J.-P. et al. (1993) Biochem. J. **293**, 329–337
15 Bobek, L. A., Tsai, H., Biesbrock, A. R. and Levine, M. J. (1993) J. Biol. Chem. **268**, 20563–20569
16 Gum, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1994) J. Biol. Chem. **269**, 2440–2446
17 Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M. and Rose, M. C. (1994) J. Biol. Chem. **269**, 12932–12939
18 Gum, J. R., Hicks, J. W., Lagace, R. E. et al. (1991) J. Biol. Chem. **266**, 22733–22738
19 Xu, G., Huan, L.-J., Khatri, I. A. et al. (1992) J. Biol. Chem. **267**, 5401–5407
20 Huan, L. J., Xu, G., Forstner, G. and Forstner, J. (1992) Biochim. Biophys. Acta **1132**, 79–82
21 Bhargava, A. K., Woitach, J. T., Davidson, E. A. and Bhavanandan, V. P. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 6798–6802
22 Eckhardt, A. E., Timpte, C S., Abernethy, J. L., Zhao, Y. and Hill, R. L. (1991) J. Biol. Chem. **266**, 9678–9686
23 Shankar, V., Gilmore, M. S. and Sachdev, G. P. (1992) Biochem. Biophys. Res. Commun. **189**, 958–964
24 Verma, M. and Davidson, E. A. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 7144–7148
25 Hoffman, W. (1988) J. Biol. Chem. **263**, 7686–7690
26 Probst, J. C., Gertzen, E.-M. and Hoffman, W. (1990) Biochemistry **29**, 6240–6244
27 Hauser, F. and Hoffman, W. (1992) J. Biol. Chem. **267**, 24620–24624
28 Afdhal, N. H. and Smith, B. F. (1990) Hepatology **11**, 699–702
29 Klinkspoor, J. H., Tytgat, G. N. J. and Groen, A. K. (1993) Eur. J. Gastro. Hepatol. **5**, 226–234
30 Smith, B. F. (1987) J. Lipid Res. **28**, 1088–1097
31 Carey, M. C. and Cahalane, M. J. (1988) Gastroenterology **95**, 508–523
32 Lee, S. P., LaMont, J. T. and Carey, M. C. (1981) J. Clin. Invest. **67**, 1712–1719
33 Lee, S. P., Carey, M. C. and LaMont, J. T. (1981) Science **211**, 1429–1432
34 Afdhal, N. H., Offner, G. D. and Smith, B. F. (1990) Gastroenterology **99**, 1493–1501
35 Chomczynski, P. and Sacchi, N. (1987) Anal. Biochem. **162**, 156–159
36 Feinberg, A. P. and Vogelstein, B. (1984) Anal. Biochem. **137**, 266–267
37 Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
38 Wijngaard, P. L. J., Metzelaar, M. J., MacHugh, N. D., Morrison, W. I. and Clevers, H. C. (1992) J. Immunol. **149**, 3273–3277
39 Kodama, T., Freeman, M., Rohrer, L., Zabrecky, J., Matsudaira, P. and Krieger, M. (1990) Nature (London) **343**, 531–535
40 Jones, N. H., Clabby, M. L., Dialynas, D. P., Huang, H.-J. S., Herzenberg, L. A. and Strominger, J. L. (1986) Nature (London) **323**, 346–349
41 Aruffo, A., Melnick, M. B., Linsley, P. S. and Seed, B. (1991) J. Exp. Med. **174**, 949–952
42 Dangott, L. J., Jordan, J. E., Bellet, R. A. and Garbers, D. L. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 2128–2132
43 Journet, A. and Tosi, M. (1986) Biochem. J. **240**, 783–787
44 Leytus, S. P., Kurachi, K., Sakariassen, K. S. and Davie, E. W. (1986) Biochemistry **25**, 4855–4863
45 Shimell, M. J., Ferguson, E. L., Childs, S. R. and O'Connor, M. B. (1991) Cell **67**, 469–481
46 Freeman, M., Ashkenas, J., Rees, D. J. G. et al. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 8810–8814
47 Chou, P. Y. and Fasman, G. D. (1979) Biophys. J. **26**, 367–383
48 Kyte, J. and Doolittle, R. F. (1982) J. Mol. Biol. **157**, 105–132
49 Kim, Y. S., Gum, J. R., Byrd, J. C. and Toribara, N. W. (1991) Am. Rev. Respir. Dis. **144**, S10–S14
50 Gum, J. R. (1992) Am. J. Respir. Cell Mol. Biol. **7**, 557–564
51 Shelton-Inloes, B. B., Titani, K. and Sadler, J. E. (1986) Biochemistry **25**, 3164–3171
52 Bell, A. E., Sellers, L. A., Allen, A., Cunliffe, W. J., Morris, E. R. and Ross-Murphy, S. B. (1985) Gastroenterology **88**, 269–280
53 Sellers, L. A., Allen, A., Morris, E. R. and Ross-Murphy, S. B. (1988) Carbohydr. Res. **178**, 93–110
54 Van de Velde, H., von Hoegan, I., Luo, W., Parnes, J. R. and Thielemans, K. (1991) Nature (London) **351**, 662–665
55 Afdhal, N. H., Offner, G. D., Murray, F. E., Troxler, R. F. and Smith, B. F. (1990) Gastroenterology **98**, 1633–1641
56 Smith, B. F. and Lamont, J. T. (1984) J. Biol. Chem. **259**, 12170–12177

Received 10 October 1994/22 February 1995; accepted 5 April 1995