Bovine submaxillary mucin contains multiple domains and tandemly repeated non-identical sequences

Weiping JIANG[1], Joseph T. WOITACH[2], Ralph L. KEIL and Veer P. BHAVANANDAN

Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA 17033, U.S.A.

A number of cDNA fragments coding for bovine submaxillary mucin (BSM) were cloned, and the nucleotide sequence of the largest clone, BSM421, was determined. Two peptide sequences determined from the purified apoBSM were found near the N-terminus of the mucin-coding region of BSM421. This clone does not contain a start or stop codon, but its 3′ end overlaps with the 5′ end of a previously isolated clone, λBSM10. The composite sequence of 1589 amino acid residues consists of five distinct protein domains, which are numbered from the C-terminus. The cysteine-rich domain I can be further divided into a von Willebrand factor type C repeat and a cystine knot. Domains III and V consist of similar repeated peptide sequences with an average of 47 residues. Domains II and IV do not contain such sequences but are similar to domains III and V in being rich in serine and threonine, many of which are predicted to be potential O-glycosylation sites. Domain III also contains two sequences that match the ATP/GTP-binding site motif A (P-loop). Only β-strands and no α-helices are predicted for the partial deduced amino acid sequence. Northern analysis of submaxillary gland RNA with the BSM421 probe detected multiple messages of BSM with sizes from 1.1 to over 10 kb. The tandemly repeated, non-identical peptide sequences of approx. 47 residues in domains III and V of BSM differ from the tandemly repeated, identical 81-residue sequences of pig submaxillary mucin (PSM), although both BSM and PSM contain similar C-terminal domains. In contrast, two peptide sequences of ovine submaxillary mucin are highly similar (86% and 65% identical respectively) to the corresponding sequences in domain V of BSM.

INTRODUCTION

Mucins are the major macromolecular constituent of the mucous secretions that coat the oral cavity and the respiratory, gastrointestinal and urogenital tracts of animals [1–5]. They are responsible for the viscoelastic properties of the secretions, providing protection for the exposed delicate epithelial surfaces from microbial and physical insults [5–7]. Secretory mucins are typically of very high molecular mass (over 1 MDa) and have hundreds of O-linked saccharides, constituting between 50% and 80% of the molecule by weight. The saccharides are based, at present, on seven core structures and can vary in length from disaccharides to large oligosaccharides of approx. 20 monosaccharides and exhibit astonishing diversity [5,8]. The biological relevance of this diversity is not fully understood, but one possibility is that they act as 'decoy' receptors for the prevention of the binding of pathogens to epithelial cells [5,9]. It has been known for a long time that the saccharides are linked to serine and threonine residues of the protein scaffold. However, owing to the technical problems associated with deglycosylation of mucins, the biochemical characterization of the protein backbone of the large secreted mucins has been fraught with difficulties [10,11]. The application of cDNA cloning has improved the situation and provided partial amino acid sequences of several epithelial mucins that reveal new features and complexities such as the presence of tandem repeats and cysteine-rich domains. Considering only the large-molecular-mass epithelial secreted mucins, so far the complete amino acid sequences of only the human intestinal/tracheal MUC2 mucin and the frog integumentary mucin B.1 have been deduced from the cDNA sequences [12,13]. However, it is important to point out that neither of the sequences has been confirmed by direct amino acid sequence analysis. Partial deduced amino acid sequences are available for several epithelial mucins, including pig submaxillary mucin (PSM) [14,15], human intestinal acidic mucin (MUC3) [16], human tracheobronchial mucins [17,18], and human stomach mucin (MUC6) [19]. Of these, only the sequence of PSM has been confirmed by the direct sequence analysis of peptide fragments [14].

We have been investigating the polymeric structure of bovine submaxillary mucin (BSM), which has properties very similar to those of its sheep counterpart (OSM). Previously the screening of a bovine submaxillary gland λgt11 cDNA library with anti-apoBSM antibodies yielded a 2.0 kb clone (λBSM10) [20]. Because the amino acid composition of the deduced sequence did not match that of BSM, it was presumed to code for an unidentified mucin-like protein [20]. The most interesting feature of the cloned protein was the presence of two distinct domains, one being a cysteine-rich C-terminal domain that has subsequently been shown to be present in a number of other epithelial mucins [12,15,21–23]. For example, there is 82% similarity, including 30 identical half-cystine residues, between the C-terminal sequences of the λBSM10 clone and PSM [15]. Half-cystine residues in the cysteine-rich domain of BSM and PSM can also be aligned with those in the corresponding domain of human pre-pro-von Willebrand factor (vWF), although the overall sequence identity is low, e.g. 20% between PSM and vWF [15].

Here we present a partial amino acid sequence of BSM deduced from the nucleotide sequence of BSM421, the largest clone isolated from a λZAPII DNA library. BSM421 contains a

Abbreviations used: PSM, BSM and OSM, pig, bovine and sheep submaxillary mucin; CA VI, carbonic anhydrase VI; vWF, von Willebrand factor.

[1] To whom correspondence should be addressed.

[2] Present address: Laboratory of Experimental Carcinogenesis, National Cancer Institute, National Institutes of Health, Building 37, Room 3C04, Bethesda, MD 20892, USA.

The nucleotide sequence reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence. Databases under the accession number AF016589.

3.6 kb nucleotide sequence that overlaps with the 5′ end of the λBSM10 sequence. Two peptide sequences determined for apo-BSM were found near the N-terminus of BSM421. There was no significant sequence similarity to PSM outside the C-terminal domains mentioned above.

EXPERIMENTAL

Screening of a cDNA library

Construction of the bovine submaxillary cDNA library in λZAPII was as described previously [24]. The library was plated on NZY plates at a dilution that provided 5000–10000 plaque-forming units per plate. Duplicated filter lifts were performed on the plates by using Magna nylon transfer membranes (Micron Separations). Filters were washed, processed and exposed to a UV transilluminator for 5 min [25]. They were prehybridized for at least 2 h at 65 °C in a solution containing 6 × SSC (0.9 M NaCl/0.09 M sodium citrate), 5 × Denhardt's solution (0.1 % polyvinylpyrrolidone/0.1 % BSA/0.1 % Ficoll 400), 0.2 % SDS and 200 μg/ml fragmented, denatured salmon sperm DNA. They were then hybridized at 65 °C for 16 h in fresh pre-hybridization buffer containing the radiolabelled probe at 5×10^6 c.p.m./ml. Filters were washed to a final stringency of 0.5 × SSC/0.1 % SDS at 65 °C for 15 min, then exposed to X-ray film. Putative plaques were verified by secondary screening. Positive clones were isolated and their inserts were subcloned into pBluescript SK − by excision *in vivo* with the helper phage R408 (Stratagene). Plasmids were isolated with the QIAprep Spin Miniprep Kit (Qiagen).

Northern blot analysis

Total RNA was also isolated from various bovine tissues with TRI-Reagents (Sigma). To remove potential contaminations with DNA and protein, the RNA preparations were either further extracted with acidic phenol or treated with RNase-free DNase I [26]. Total RNA was electrophoresed on 1 % (w/v) agarose/2.2 M formaldehyde denaturing gels and transferred to nylon membranes. The membranes were hybridized with DNA probes at 42 °C overnight in a solution containing 6 × SSC, 5 × Denhardt's solution, 30 % (v/v) formamide and 0.1 mg/ml denatured herring sperm DNA. After hybridization the blots were washed with 2 × SSC containing 0.1 % SDS, first at 25 °C, then at 42 °C and finally at 55 °C. Blots were air-dried and autoradiographed with Kodak XAR-5 film at −70 °C.

DNA sequence analysis

Double-stranded plasmid DNA was sequenced either manually with Sequenase (version 2.0) (Amersham) or automatically with AmpliTaq DNA polymerase, using an ABI Prism automated DNA sequencer in the Macromolecular Core Facility at the Pennsylvania State University College of Medicine. The ends of clones were sequenced with the primers derived from vectors such as T3, T7 and M13 reverse primer. The internal regions were sequenced with gene-specific primers based on determined sequences or by generating deletion subclones with restriction enzymes.

DNA and protein databases were searched with determined DNA sequences and deduced amino acid sequences by using NCBI's BLASTN and BLASTP with BCM's BEAUTY post-processing, which adds annotated domain information [27]. Similar sequences were aligned with CLUSTAL W [28]. Potential secondary structure elements were predicted by NNSSP and PHD [29,30]. The above computing was performed with the BCM Search Launcher [31] at the Human Genome Center (Baylor College of Medicine, Houston, TX, U.S.A.). The ExPASy server [32] from the Geneva University Hospital and University of Geneva (Geneva, Switzerland) was used to predict potential O-glycosylation sites with NETOGLYC 2.0 [33] and to search Prosite with SCANPROSITE and PROFILESCAN [34,35].

RESULTS

Cloning and sequencing of BSM421

The λZAPII cDNA library was constructed from poly(A)$^+$ RNA isolated from the submaxillary gland of a single cow [24]. The library contained 3.5×10^6 phages, of which 98 % were recombinants. After screening of 1.5×10^6 recombinant phages with the λBSM10 DNA fragment as a probe, over 100 positive clones were detected, indicating a large abundance of mRNA species related to λBSM10. To identify clones that further extended the 5′ end of λBSM10, the positive clones were rescreened with an oligonucleotide to the 5′ end of λBSM10. The 40 clones detected by this oligonucleotide were plaque-purified, and their inserts were subcloned into pBluescript SK − by excision *in vivo* with the helper phage R408 (Stratagene). The cDNA insert sizes of these clones ranged from 2.0 to 2.9 kb, as determined by restriction endonuclease digestion. The 2.9 kb clone was partly sequenced, and an oligonucleotide of the 5′ end of the 2.9 kb clone was used to rescreen the cDNA library. Screening of 2×10^5 recombinant phages yielded 20 clones with cDNA insert sizes ranging from 2.8 to 4.6 kb. The cDNA species of these 20 clones were mapped with restriction enzymes and portions sequenced. Both strands of the largest clone, BSM421, were completely sequenced (Figure 1).

Nucleotide and amino acid sequences of BSM421 and λBSM10

BSM421 was found to be a hybrid clone containing two open reading frames in opposite orientations (Figure 1). The reading

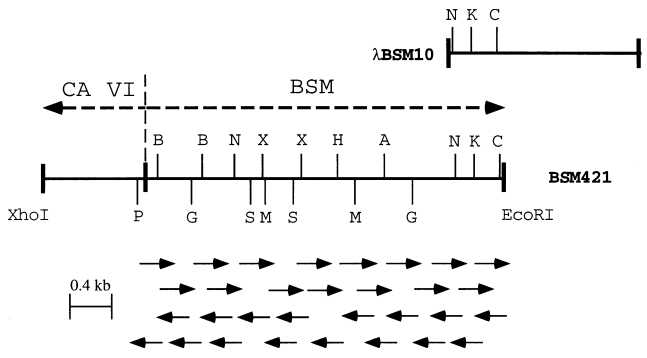

Figure 1 Restriction map and sequencing strategy of BSM421

The horizontal line represents the 4.6 kb insert flanked by *Eco*RI and *Xho*I sites contained in the linker sequences used in the library construction. The 1 kb sequence near the *Xho*I site was found to correspond to the C-terminal portion and the poly(A)$^+$ tail of CA VI [24]. The 3.6 kb sequence near the *Eco*RI site codes for mucin. The two open reading frames are in opposite orientations, shown as broken lines, with arrows indicating the direction of the reading frames. The relationship of BSM421 to λBSM10, a clone previously isolated from a different library [20], is shown. Restriction enzymes shown above and below the horizontal line were used to construct subclones for sequencing. Primers (T3 and T7) for the vector, pBluescript SK −, and synthetic oligonucleotides were used as primers for sequencing reactions. Solid arrows indicate the extent and direction of the sequence information obtained for the respective subclones. Restriction enzymes are: *Sph*I (P), *Bam*HI (B), *Bgl*II (G), *Nde*I (N), *Spe*I (S), *Xba*I (X), *Mun*I (M), *Hpa*I (H), *Acc*I (A), *Kpn*I (K) and *Nco*I (C). The scale of the drawing is indicated.

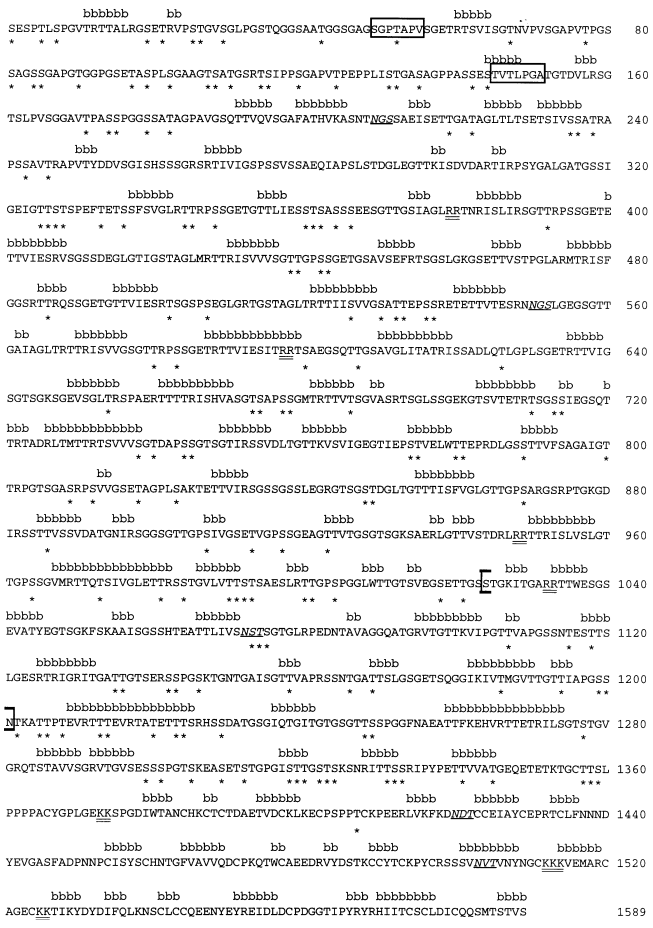

Figure 2 Deduced amino acid sequence of BSM421 and λBSM10

The amino acid sequence, numbered at the right, is deduced from the composite nucleotide sequence of BSM421 and λBSM10 with the overlapping regions indicated by square brackets. Boxed sequences are identical with peptide sequences determined from the purified apoBSM. Ser and Thr residues indicated by asterisks are potential O-glycosylation sites predicted by the NETOGLYC 2.0 program [33]. Tripeptides underlined and in italics are potential N-glycosylation sites. Dibasic and tribasic peptides are doubly underlined. Lower-case letters 'b' above the sequence represent β-strand secondary structure predicted by NNSSP [29]. No α-helices are predicted.

frame near the *Xho*I site encodes the C-terminal portion of carbonic anhydrase (CA) VI followed by a poly(A)$^+$ tail [24]. The reading frame near the *Eco*RI site codes for the mucin. As stated previously, how this fusion occurred is not known, although the restriction enzymes (*Eco*RI and *Xho*I) used in the library construction were not involved and fusion of the CA VI and mucin represented a rare event in the library [24]. The mucin-coding region of BSM421 consists of 3603 nt encoding 1201 amino acid residues (Figure 2). Two peptide sequences near the N-terminal region of the mucin-coding region from BSM421 are identical with those (I-7-1 and I-11) determined from the deglycosylated purified BSM (V. P. Bhavanandan, M. Sheykhnazari, J. T. Woitach and K. Yamamoto, unpublished work). This finding provides direct evidence that BSM421 encodes a portion of the BSM protein.

The 3′ end of BSM421 overlaps with the 5′ end of λBSM10, which was previously thought to encode a full-length mucin-like protein because it contained a start codon at its 5′ end and the amino acid composition differed from that of the purified mucin [20]. However, there were nucleotide differences between the two sequences, especially in the region of nt 3070–3165. The reason for these discrepancies could be allelic differences, because BSM421 and λBSM10 were isolated from libraries constructed from two different cows. However, resequencing of λBSM10 revealed that these discrepancies were due to errors introduced into the published λBSM10 sequence. The following corrections were made for the λBSM10 cDNA sequence, and nucleotide numbering is the same as that used for the published sequence [20]. Nucleotides T and C at positions 28 and 29 were replaced with a single nucleotide G, nucleotide G at position 32 was replaced by A, and nucleotide C at position 104 was deleted. These corrections led to changes in the reading frames and extended the deduced amino acid sequence to the 5′ end without a start codon.

The above changes indicate that λBSM10 is not a full-length clone as previously shown [20]. However, the major conclusion derived from the published λBSM10 sequence still holds. That is, BSM contains two distinct domains near its C-terminus, one of which is the cysteine-rich region now found in many mucin or mucin-like proteins [20]. The revised λBSM10 sequence is identical with the BSM421 sequence in the 534 bp region (corresponding to amino acid residues 1024–1201 in Figure 2) and extends the latter to the poly(A)$^+$ tail.

Domains of BSM

The composite sequence of 1589 amino acid residues of BSM421 and λBSM10 clones is divided into five domains on the basis of unique features present in each region (Figure 3). Domains are numbered from the C-terminus because the N-terminal regions have not been cloned. Domain I is the 234-residue C-terminal region (positions 1356–1589) containing 30 cysteine residues as identified previously [20]. Two sequences in this domain match two signature sequences of Prosite [34]. One is used to detect vWFc domains, which are named after the vWF type C repeat [36]. The other signature sequence identifies a cystine knot that is found in a structural superfamily of growth factors, including transforming growth factor β and nerve growth factor [37]. Therefore domain I can be further divided into two cysteine-rich subdomains, a vWFc domain and cystine knot.

In contrast, domain II (residues 1048–1355) is rich in serine and threonine residues, as previously identified [20] (Figure 3). It contains an 11-residue sequence that is present three times, the first two of which are followed by a pentapeptide [20]. In addition, six peptide segments can be aligned to each other on the basis of two common threonine residues shown in Figure 3 in larger bold letters and marked by a #. However, the overall similarity is low and might not be significant.

In comparison, domain III contains 15 peptide segments that are more similar to each other (Figure 3). On the basis of the length and similarity of peptide segments, domain III is further divided into two regions, IIIa and IIIb. Seven segments of slightly different lengths in region IIIa (residues 720–1047) are more similar to each other than those in domain II: 42% of the residues (21 of 50), shown in bold letters and marked by ^, are conserved in more than half of the sequences in region IIa. Several gaps are, however, required for the optimal alignment in region IIIa. Eight segments of 47 residues in region IIIb (376 residues from positions 344–719) are more similar to each other than those in region IIIa. In region IIIb, no insertions or deletions are present in the aligned sequences in which two serine and two threonine residues are conserved in all eight sequences and a further 32 residues are conserved (68%) in at least five of the eight sequences. In addition, domain III contains two identical

```
(V)
  1                                  SESPTLSPGVTRTTALRGSE    21
 21  TRVPSTGVSGLPGSTQGGSAATGGSGAGSGPTAPVSGETRTSVI-SG    66
 67  TNVPVSGAPVTPGSSAGSSGAPGTGGPGSETASPLSGAAGTSATGSR   113
114  TSIPPSGAPVTPEPPLISTGASAGPPASSESTVTLPGATGTDVLRSG   160
161  TSLPVSGGAVTPASSPGGSSATAGPAVGSQTTVQVSGAFATHVKASN   207
     #  # ^#  ^^# ^  ^ ^ #  ^   ^^  ^   ^^^  # ^  #
(IV)
208                      TNGSSAEISETTGATAGLTLTSETSIVSSA    237
238  TRAPSSAVTRAPVTYDDVSGISHSSSGRSRTIVIGSP             274
275            SSVSSAEQIAPSLSTDGLEGTTKISDVDAR            304
305  TIRPSYGALGATGSSIGEIGTTSTSPEFTETSSFSVGLR           343
                        ^ ^         ^
(IIIb)
344  TTRPSSGETGTTLIESSTSASSSEESGTTGSIAGLRRTNRISLIRSG   390
391  TTRPSSGETETTVIESRVSGSSDEGLGTIGSTAGLMRTTRISVVVSG   437
438  TTGPSSGETGSAVSEFRTSGSLGKGSETTVSTPGLARMTRISFGGSR   484
485  TTRQSSGETGTTVIESRTSGSPSEGLGRTGSTAGLTRTTIISVVGSA   531
532  TTEPSSRETETTVTESRNNGSLGEGSGTTGAIAGLTRTTRISVVGSG   578
579  TTRPSSGETRTTVIESITRRTSAEGSQTTGSAVGLITATRISSADLQ   625
626  TLGPLSGETRTTVIGSGTSGKSGEVSGLTRSPAERTTTTRISHVASG   672
673  TSAPSSGMTRTTVTSGVASRTSGLSSGEKGTSVTETRTSGSSIEGSQ   719
     #^ ^^#^^# ^^^^^^ ^^^ ^ ^^^^^^^^ ^^^ ^^^^^#    ^
(IIIa)
720  TTRTADRLTMTTTRTSVVVSGTDAPSSGTSGTIRSSVDLTGTTKVSVIGEG  769
770  TIEPSTVELWTTEPRDLGSSTTVFSAGAIGTTR--PGTSGASRPSVVGSE  827
828  TAGPLSAKTETTTVIRSGSSGSSLEGRGTSGST---DGLTGTTTISFVGLG  864
865  TTGPSARGSRPTGKGDIRSSTTVSSVDATGNIRS--GGSGTTGPSIVGSE  902
903  TVGPSSGEAGTTVTGSGTSGKSAERLGTTVST---DRLRRTTRISLVSLG  959
960  TTGPSSGVMRTTQTSIVGLETTRSSTGVLVTT---STSAESLR         999
1000 TTGPSPGGLWTTGTSVEGSETTGSSTGKITGAR--RTTWESGSEVATYEG 1047
     # ^^^     ^#      ^ ^^  ^ ^  ^ ^^        ^^  ^ ^  ^
(II)
1048 TSGKFSKAAISGSSHTEATTLIVSNSTSGTGLRPEDNTAVAGGQATGRVTGT  1099
1100 TKVIPGTTVAPGSSNTESTTSLGESRTRIGRITGATTGTSERSSPGSKTGN  1150
1151 TGAISGTTVAPRSSNTGATTSLGSGETSQGGIKIVTMGVTTGTTIAPGSSNTK 1203
1204 TTPTEVRTTTEVRTATETTTSRHSSDATGSGIQTGITGTGSGTTSSPGGFNAEA1258
1259 TTFKEHVRTTETRILSGTSTGVGRQTSTAVVSGRVTGVSESSSPGTSKEASE  1310
1311 TSTGPGISTTGSTSKSNRITTSSRIPYPETTVVATGEQETETKTG        1355
     #           ^ ^  ^#
(I)                                  VWFC domain signature
1356 CTTSLPPPPACYGPLGEKKSPGDIWTANCHKCTCTDAETVDCKLKECPSPPTC 1408

1409 KPEERLVKFKDNDTCCEIAYCEPRTCLFNNNDYEVGASFADPNNPCISYSCHN 1461
1462 TGFVAVVQDCPKQTWCAEEDRVYDSTKCCYTCKPYCRSSSVNVTVNYNGCKKK 1514
                                     C-terminal cystine
1515 VEMARCAGECKKTIKYDYDIFQLKNSCLCCQEENYEYREIDLDCPDGGTIPYR 1567
     knot signature
1568 YRHIITCSCLDICQQSMTSTVS                               1589
```

Figure 3 Domains of BSM

Five domains (I to V, numbered from the C-terminus) with unique features are present in the deduced partial amino acid sequence of BSM. The numbering of amino acid residues, shown at the left and right of the sequence, is the same as that used in Figure 2. The sequences overlined in domain I match the signature sequences for the vWFc domain and the cystine knot [37,40]. Sequences repeated in domain II are underlined. Segments in domains II to V are compared with each other. Domain III is further divided into IIIa and IIIb. For each region, the residues identical in all peptide sequences are shown in larger bold letters and labelled with a #, and the residues conserved in more than half of peptide sequences are in bold letters and labelled with a ^. Dashes are gaps introduced for the optimal alignment. Sequences overlined in domain III are identical and match the consensus sequence of ATP/GTP-binding site motif A [38].

sequences that match the ATP/GTP-binding site motif A (P-loop) that is conserved in many proteins that bind ATP or GTP [38].

Tandemly repeated, non-identical sequences of domain III are not present in domain IV (residues 208–343), which contains peptide segments that do not show significant similarity to each other (Figure 3). However, similar repeated sequences are present in domain V (residues 1–207). One gap in domain V is required for optimal alignment, and 42% (20 of 47) of the residues are conserved in more than half of the sequences.

Predicted glycosylation sites and secondary structure of BSM

There are many serine residues (82 of 269; 30.5%) and threonine residues (89 of 298; 29.9%) that are predicted as potential O-glycosylation sites by NETOGLYC 2.0 [33] (Figure 2). Most of these sites are within domains II to V, which are rich in serine and threonine residues, although four potential O-glycosylation sites are also present in the cysteine-rich domain I. There are five tripeptides that match the consensus sequence of Asn-Xaa-Ser/Thr for potential N-glycosylation (Figure 2). Three of the O-glycosylation sites overlap with one of the N-glycosylation sites. However, for some proteins such as PSM it has been shown that the extent of O-glycosylation of serine and threonine residues is poorly correlated with published predictive methods [39]. Work is under way to determine O-glycosylation sites for BSM.

The secondary structure of BSM is predicted by several programs, including NNSSP and PHD [29,30]. The results from these predictions are similar; the NNSSP results are shown in Figure 2. It is very interesting that, throughout the entire cloned polypeptide chain, only β-strands and no α-helices or other structures are predicted. It is surprising that the predicted secondary structure for domain I is similar to that of the other four domains (II to V), considering that the sequence and the amino acid composition of domain I are very different from those of domains II to V.

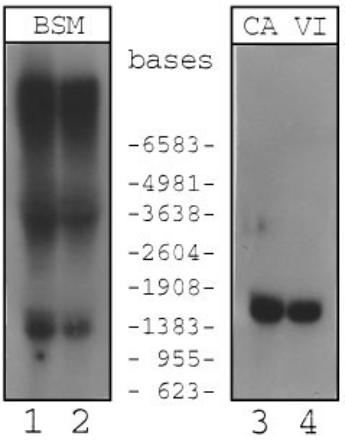

Figure 4 Northern blot analyses of bovine submaxillary gland RNA

Total RNA was isolated with TRI-Reagents and extracted further by the acidic phenol procedure to remove DNA and protein [26]. Results of two RNA preparations from different cows analysed by Northern blots are illustrated. Samples in lanes 1 and 2 were probed with a 2.2 kb *Bg*II–*Bg*II fragment derived from BSM421. The blots were stripped and probed again with a CA VI cDNA (lanes 3 and 4) [24]. Sizes of RNA standards are indicated.

Multiple BSM messages in the submaxillary gland

The tissue distribution of the BSM mRNA was examined by Northern blot analyses with probes generated from BSM421. The BSM messages are highly specific to the submaxillary gland and were not detected in any other bovine tissues tested, including mucin-secreting organs such as lung, stomach, small intestine and large intestine, as well as brain, kidney, heart and liver [20]. Sizes of the BSM mRNA in the submaxillary gland ranged from 1.0 kb to more than 10 kb (Figure 4, lanes 1 and 2). Several signals, including 1.1 kb (1.3 kb, minor), 3.3 kb (3.9 kb, minor), and more than 10 kb, were discrete and distinct from two smear regions around 3–5 kb and above 7 kb. The signals larger than 5 kb most probably represent the messages of BSM, because the size of the cloned cDNA is approx. 5 kb and is still incomplete at its 5′ end. The signals smaller than 5 kb and the smear region might represent the messages derived from highly similar genes and/or alternative splicing of the same RNA precursors. However, the possibility that the smear regions were due to degradation of the very large transcripts cannot be excluded, although a single 1.4 kb mRNA species was detected when the same samples were probed again with the bovine CA VI cDNA (Figure 4, lanes 3 and 4).

DISCUSSION

Primary structure of BSM

Cloning and sequencing of λBSM10 revealed for the first time the presence of a cysteine-rich domain in the C-terminal region of mucin-like proteins [20]. The 5′ end of λBSM10 cDNA does not, however, contain the initiation start codon, as originally described. The corrected nucleotide sequence of λBSM10 (1.9 kb) has been extended by approx. 3 kb by cloning and sequencing BSM421. The composite sequence of BSM421 and λBSM10 consists of a total of approx. 5 kb of cDNA sequence encoding almost 1600 amino acid residues. Most importantly, the deduced amino acid sequence of BSM421 contains peptide sequences determined for the BSM protein, showing for the first time that the cDNA clones (BSM421 and λBSM10) encode portions of BSM. However, the composite cDNA sequence is still incomplete at its 5′ end, because it lacks an initiation codon for translation.

A total of five domains with unique sequence features are present in the deduced amino acid sequence of BSM (Figure 2). The cysteine-rich C-terminal domain I (234 residues) can be divided further into the vWFc domain and the cystine knot on the basis of the presence of these signature sequences in this domain. The presence of the cystine knot in BSM is further supported by the prediction of a β-strand structure for this region, and by the fact that members of the cystine knot family are known to contain a conserved β-strand structure [37]. The cystine knot has been postulated to be a dimerization domain [36]. The vWFc domain, duplicated in the vWF, is thought to participate in oligomerization but not in the initial dimerization step [40]. The oligomerization status of BSM has been difficult to determine by direct analysis of the native mucin, because of its very high molecular mass and high carbohydrate content. On the assumption that it forms oligomers, it is possible that the cystine knot of BSM is involved in the initial dimerization step and that the vWFc domain of BSM participates in subsequent oligomerization steps to form larger protein complexes. This might be true in general for several other epithelial mucins containing these two domains. In fact, the cysteine-rich domain of BSM is very similar to that of PSM, which, as described below, has been shown to form a disulphide-linked dimer [41]. Similarly the human MUC2 mucin apoprotein containing the cysteine-rich domain seems to dimerize [12,42,43].

Domains III and V of BSM consist of tandemly repeated similar sequences of an average of 47 residues. Therefore BSM is similar to most mucin core proteins so far sequenced, which have tandemly repeated amino acid sequences [2–4,8]. However, the 47-residue repeated sequences deduced from the BSM cDNA differ from repetitive sequences of 28 residues predicted for BSM by Pigman et al. [44] on the basis of amino acid composition analysis of tryptic glycopeptides. The possible 28-residue repeats could be in uncloned N-terminal portions of the same BSM gene or be encoded by a different gene.

In addition, domain III of BSM contains two identical peptides (GSGTSGKS) that match the consensus sequence of the ATP/GTP-binding site motif (P-loop) that has been found in several protein classes, such as kinases, transporters and structural proteins [38]. To our knowledge BSM is the first mucin protein that has been identified to contain putative ATP/GTP-binding sequences. The possibility that Ser and Thr residues in these sequences of BSM are glycosylated cannot be excluded, although none of these amino acids is predicted by NETOGLYC 2.0 to be glycosylated (Figure 2). Whether BSM binds ATP and/or GTP *in vivo* is not known. This would be an important area to explore, considering the potentially important consequences of nucleotide binding on a mucin molecule. For example, ATP is known to stimulate mucin exocytosis [45]. In addition, ATP and/or GTP binding could lead to enzymic activity or to changes in characteristics, such as charge and conformation, that affect interactions with other molecules.

The above domains of BSM predicted from the deduced amino acid sequence might not all be present in the mature core protein if the biosynthesis of BSM involves post-translational processing. Such processing for BSM is likely, because the size of the mature apoBSM is considerably smaller than would be predicted from the deduced sequences [46]. There are six dibasic peptides (Arg-Arg or Lys-Lys) and one tribasic peptide (Lys-Lys-Lys) that could be sites for potential post-translational processing by trypsin-like enzymes (Figure 2).

Multiple messages of BSM

Polydisperse messages have been observed for transcripts of genes encoding mucins, including BSM and PSM [15,17,18,20,47,48]. The smear regions in the Northern blots (Figure 4) suggest heterogeneity (polydispersity) of BSM messages. However, the relatively discrete and strong signals of 1.1, 3.3 and over 10 kb in the same blots indicate that the sizes of certain BSM messages are unique and well defined, and that these messages are more abundant than others. Distinct bands have also been observed for the transcripts of other mucin genes such as rat intestinal mucins and human *MUC2*, *MUC4* and *MUC5AC* [23,49–51]. In addition, the same RNA preparations produce different hybridization patterns when different probes derived from the same BSM cDNA are used (results not shown). The complex hybridization patterns of the BSM transcripts suggest the presence of multiple copies of similar genes in the bovine genome and/or alternative splicing of same mRNA precursors. Preliminary results from analysis of the cloned bovine genome DNA fragments provide evidence supporting both possibilities (W. Jiang, D. Gupta, J. T. Woitach and V. P. Bhavanandan, unpublished work).

Comparison of BSM and OSM

BSM has saccharide structures and biochemical properties similar to those of OSM [52,53], which has not been cloned. We have determined one peptide sequence (Rt25.5) from the purified OSM that showed 86% identity with a peptide in the deduced amino acid sequence of BSM (Figure 5A). In addition, Hill et al. [10] determined sequences of three tryptic peptides derived from OSM. A region (residues 17–50) from one of these peptides (OSM T1) is 65% identical with the N-terminal sequence (residues 1–34) of the cloned BSM (Figure 5A). The high similarity between BSM and OSM peptide sequences suggests for the first time that these two proteins are also similar in primary structure.

There is one amino acid difference in the overlapping region between OSM Rt25.5 and OSM T1, suggesting that these two OSM peptides might not be derived from a continuous sequence and that OSM might contain similar sequences in different regions. Residues 1–16 of OSM T1 and the other two OSM peptides are not found in the cloned BSM cDNA. It is possible that these peptides are unique to OSM or that they are N-terminal to the deduced BSM sequence. Extension of the OSM T1 sequence beyond the 5′ end of the BSM421 sequence favours the second possibility.

Domain structure of BSM and PSM

DNA and protein database searches with the BSM sequences shows that the C-terminal domains of BSM (I and II) are highly

Figure 5 Comparison of the BSM sequence with that of OSM and PSM

(**A**) OSM Rt25.5 was determined from the purified OSM. OSM T1 was reported by Hill et al. [10]. Identical residues between the BSM sequence (1–42) and the OSM peptides are indicated by vertical lines. (**B**) The composite sequence of BSM and the sequence of PSM103 were aligned by CLUSTAL W [28]. The BSM sequence starting at residue 294 is numbered the same as in Figure 2. The PSM sequence starting at residue 1 is numbered in accordance with the published sequence [15]. The identical residues between BSM and PSM are indicated by asterisks. The first residues of tandemly repeated sequences (threonines of the 47-residue repeats of BSM and glycines of the 81-residue repeats of PSM) are shown in small bold letters and underlined. Boundaries for domains I to IV of the BSM sequence as described in Figure 3 are indicated by roman numerals. In domain I, 30 cysteine residues conserved in BSM and PSM are shown in bold letters.

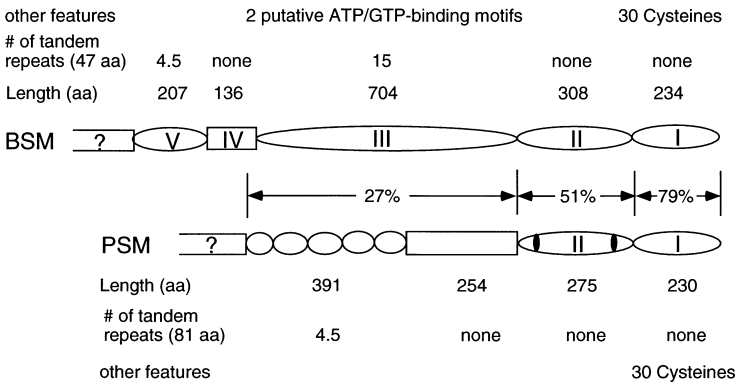

Figure 6 Domain structure of BSM and PSM

BSM domains are described in Figure 3. PSM domains are based on sequence comparison, as shown in Figure 5(B). The length (amino acid residues) of each individual domain is indicated above (BSM) or below (PSM) the domain structure, as are the number of tandem repeats and other sequence features in each individual domain. The cDNA sequence of the N-terminal region, represented by an open box with a question mark, has not been determined for either BSM or PSM. Two black vertical ovals indicate two major deletions from PSM that correspond to segments of 17 and 14 residues present in domain II of BSM. Amino acid sequence identities between different regions of BSM and PSM are indicated #, number.

similar to those of PSM, as already pointed out by Eckhardt et al. [15]. Figure 5(B) shows a comparison of cloned BSM and PSM amino acid sequences, which are further illustrated in domain structures proposed for BSM and PSM in Figure 6.

Domains I of BSM and PSM are 79% identical (182 residues of 230) and do not have any deletions or insertions except for four additional residues at the C-terminus of BSM. However, the secondary structures predicted for this domain differ between BSM (only β-strands) and PSM (both helices and pleated sheets) [15]. Domain II of BSM (308 residues) and PSM (275 residues) are 51% identical (158 residues of 309 positions), with two segments of 17 and 14 residues of BSM deleted from PSM. Such deletions in PSM are interesting for the following reasons. First, three arginine residues in the 17-residue segment and two arginine residues in the 14-residue segment could be targets for trypsin-like enzymes, resulting in differential processing of the primary

translation products of BSM and PSM. Secondly, the sizes of the deletions (17 and 14 residues) in PSM are correlated with the sizes of exons determined for BSM (51 and 42 bp), suggesting that the differences between BSM and PSM could be due to the presence or absence of specific exons in the genes or to alternative splicing of similar RNA precursors in the two species (W. Jiang, D. Gupta, J. T. Woitach and V. P. Bhavanandan, unpublished work).

Although both BSM and PSM share highly similar domains (I and II) in the C-terminal portion, they contain distinct domains N-terminal to domain II. BSM contains domains III, IV and V, whereas PSM consists of a segment of 254 residues and 4.8 tandem repeats of 81 amino acid residues (391 residues). There is only 27% identity between the two sequences, and 16 segments ranging from 2 to 19 residues present in BSM are deleted from PSM. The two putative ATP/GTP-binding sequences found in BSM (residues 640–647 and 917–924) are not conserved in PSM (residues 307–314 and 544–547). Therefore it is possible that the evolution of an ancestral gene for BSM and PSM has led the more conserved C-terminal regions such as the cysteine-rich domain to serve common functions (oligomerization, for example), and the less conserved N-terminal regions to serve species-specific functions.

From analyses of genomic DNA partly digested with restriction nucleases, Eckhardt et al. [15] suggested that PSM consists of at least 25 identical repeats of the 81-residue sequence. As stated above, no such repeats are present in the N-terminal portion of BSM cloned so far. Whether BSM contains tandemly repeated identical sequences in the uncloned portion remains to be determined. To obtain DNA sequences 5′ to BSM421, cloning genomic rather than cDNA fragments would be a more practical approach, considering the large sizes of BSM messages. We are in the process of analysing several genomic fragments that might encode such sequences. The results reported here have established for the first time a partial amino acid sequence of BSM that consists of five distinct domains, two of which contain tandemly repeated non-identical sequences of an average of 47 residues.

We thank Mrs. Sharlene Washington, Dr. Krishnamoorthy Sankaran, Dr. Dwijendra Gupta and Dr. Xiaoxuan Guo for their help during the course of this work. This work was supported in part by the National Institute of Heart, Lung, and Blood Research Grant HL42651 to V.P.B.

REFERENCES

1 Carlsted, I., Sheehan, J. K., Corfield, A. P. and Gallagher, J. T. (1985) Essays Biochem. **20**, 40–76
2 Rose, M. C. (1992) Am. Physiol. J. **263**, L413–L427
3 Carraway, K. L. and Fregien, N. (1995) Trends Glycosci. Glycotech. **7**, 31–44
4 Verma, M. and Davidson, E. A. (1994) Glycoconj. J. **11**, 172–179
5 Roussel, P., Lamblin, G., Lhermitte, M., Houdret, N., Lafitte, J., Perini, J. M., Klein, A. and Scharfman, A. (1988) Biochimie **70**, 1471–1482
6 Corfield, A. P. (1992) Glycoconj. J. **9**, 217–221
7 Tabak, L. A. (1995) Annu. Rev. Physiol. **57**, 547–564
8 Bhavanandan, V. P. and Furukawa, K. (1995) in Biology of Sialic Acids (Rosenberg, A., ed.), pp. 145–196, Plenum Press, New York
9 Varki, A. (1993) Glycobiology **3**, 97–103
10 Hill, H. D., Schwyzer, M., Steinman, H. M. and Hill, R. L. (1977) J. Biol. Chem. **252**, 3799–3804
11 Bhushana Rao, K. S. P. and Masson, P. L. (1977) J. Biol. Chem. **252**, 7788–7796
12 Gum, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1994) J. Biol. Chem. **269**, 2440–2446
13 Joba, W. and Hoffmann, W. (1997) J. Biol. Chem. **272**, 1805–1810
14 Timpte, C. S., Eckhardt, A. E., Abernethy, J. L. and Hill, R. L. (1988) J. Biol. Chem. **263**, 1081–1088
15 Eckhardt, A. E., Timpte, C. S., Abernethy, J. L., Zhao, Y. and Hill, R. L. (1991) J. Biol. Chem. **266**, 9678–9686
16 Gum, J. R., Hicks, J. W., Swallow, D. M., Lagac, R. L., Byrd, J. C., Lamport, D. T. A., Siddiki, B. and Kim, Y. S. (1990) Biochem. Biophys. Res. Commun. **171**, 407–415
17 Desseyn, J.-L., Guyonnet-Dupérat, V., Porchet, N., Aubert, J.-P. and Laine, A. (1997) J. Biol. Chem. **272**, 3168–3178
18 Guyonnet-Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Galiègguue-Zouitina, S., Pigny, P., Degand, P., Aubert, J.-P. and Porchet, N. (1995) Biochem. J. **305**, 211–219
19 Toribara, N. W., Ho, S. B., Gum, E., Gum, J. R., Lau, P. and Kim, Y. S. (1997) J. Biol. Chem. **272**, 16398–16403
20 Bhargava, A. K., Woitach, J. T., Davidson, E. A. and Bhavanandan, V. P. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 6798–6802
21 Meerzaman, D., Chales, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M. and Rose, M. C. (1994) J. Biol. Chem. **269**, 12932–12939
22 Probst, J. C., Hauser, F., Joba, W. and Hoffmann, W. (1992) J. Biol. Chem. **267**, 630–6316
23 Xu, G., Huan, L., 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
24 Jiang, W., Woitach, J. T. and Gupta, D. (1996) Biochem. J. **318**, 291–296
25 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
26 Zasloff, M., Ginder, G. D. and Felsenfeld, G. (1978) Nucleic Acids Res. **5**, 1139–1152
27 Worley, K. C., Wiese, B. A. and Smith, R. F. (1995) Genome Res. **5**, 173–184
28 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) Nucleic Acids Res. **22**, 4673–4680
29 Salamov, A. A. and Solovyev, V. V. (1995) J. Mol. Biol. **247**, 11–15
30 Rost, B. and Sander, C. (1994) Proteins **19**, 55–72
31 Smith, R. F., Wiese, B. A., Wojzynski, M. K., Davison, D. B. and Worley, K. C. (1996) Genome Res. **6**, 454–462
32 Appel, R. D., Bairoch, A. and Hochstrasser, D. F. (1994) Trends Biochem. Sci. **19**, 258–260
33 Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H., Nielsen, J. O., Hansen, J.-E. S. and Brunak, S. (1995) Biochem. J. **308**, 801–813
34 Bairoch, A., Bucher, P. and Hofmann, K. (1996) Nucleic Acids Res. **24**, 189–196
35 Lüthy, R., Xenarios, I. and Bucher, P. (1994) Prot. Sci. **3**, 139–146
36 Bork, P. (1993) FEBS Lett. **327**, 125–130
37 McDonald, N. Q. and Hendrickson, W. A. (1993) Cell **73**, 421–424
38 Koonin, E. V. (1993) J. Mol. Biol. **229**, 1165–1174
39 Gerken, T. A., Owens, C. L. and Pasumarthy, M. (1997) J. Biol. Chem. **272**, 9709–9719
40 Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. and Pannekoek, H. (1991) J. Cell. Biol. **113**, 195–205
41 Perez-Vilar, J., Eckhardt, A. E. and Hill, R. L. (1996) J. Biol. Chem. **271**, 9845–9850
42 Asker, N., Baeckström, D., Axelsson, M. A. B., Carlstedt, I. and Hansson, G. C. (1995) Biochem. J. **308**, 873–880
43 Gum, J. R., Hicks, J. W., Toribara, N. W., Siddiki, B. and Kim, Y. S. (1992) J. Biol. Chem. **267**, 21375–21383
44 Pigman, W., Moschera, J., Weiss, M. and Tettamanti, G. (1973) Eur. J. Biochem. **32**, 148–154
45 Branka, J.-E., Vallette, G., Jarry, A. and Laboisse, C. L. (1997) Biochem. J. **323**, 521–524
46 Bhavanandan, V. P. and Hegarty, J. D. (1987) J. Biol. Chem. **262**, 5913–5917
47 Hoffmann, W. (1988) J. Biol. Chem. **263**, 7686–7690
48 Verma, M. and Davidson, E. A. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 7144–7148
49 Khatri, I. S., Forstner, G. G. and Forstner, J. F. (1993) Biochem. J. **284**, 391–399
50 Baeckstrom, D. and Hansson, G. C. (1996) Glycoconj. J. **13**, 833–837
51 Lesuffleur, T., Porchet, N., Aubert, J. P., Swallow, D., Gum, J. R., Kim, Y. S., Real, F. X. and Zweibaum, A. (1993) J. Cell Sci. **106**, 771–783
52 Tettamamti, G. and Pigman, W. (1968) Arch. Biochem. Biophys. **124**, 41–50
53 Wu, A. M., Csako, G. and Herp, A. (1994) Mol. Cell. Biochem. **137**, 39–55

Received 18 September 1997/17 November 1997; accepted 12 December 1997