Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6798–6802, September 1990
Biochemistry

Cloning and cDNA sequence of a bovine submaxillary gland mucin-like protein containing two distinct domains

(mucin-associated protein/cysteine-rich domain/O-glycosylation)

AJAY K. BHARGAVA*, JOSEPH T. WOITACH, EUGENE A. DAVIDSON†, AND VEER P. BHAVANANDAN‡

Department of Biological Chemistry and Program in Cell and Molecular Biology, The Milton S. Hershey Medical Center, Pennsylvania State University, Hershey, PA 17033

Communicated by Stuart Kornfeld, June 11, 1990

ABSTRACT **A λgt11 cDNA library prepared from bovine submaxillary gland mRNA was screened with polyclonal anti-apo-bovine submaxillary mucin antibodies with the aim of obtaining the deduced amino acid sequence of the mucin core protein. One of the positive clones had a 1.8 kilobase (kb) cDNA insert and coded for an incomplete protein. A 2.0-kb cDNA clone was isolated by rescreening the library with the 1.8-kb cDNA. Nucleotide sequencing of the full-length 2.0-kb cDNA revealed an open reading frame that coded for a 563-amino acid protein. A striking feature of the cloned protein is the skewed distribution of the amino acids, most notably that of the hydroxy amino acids and cysteine. The amino-terminal domain of 339 residues is very rich in threonine, serine, and glycine and poor in cysteine, aspartic acid, tyrosine, phenylalanine, and tryptophan. In contrast, the carboxyl-terminal domain of 224 residues is rich in cysteine, aspartic acid, tyrosine, lysine, and asparagine and relatively poor in threonine, serine, and glycine. A search of the protein data bank for homologies to the deduced amino acid sequence revealed statistically significant matches to several proteins, including the porcine submaxillary apomucin fragment. The cysteine-rich domain by itself was not statistically homologous with any of the registered polypeptide sequences. RNA blot analysis using DNA probes corresponding to the mucin-like and cysteine-rich regions detected a nearly identical pattern of transcripts, demonstrating that the characterized clones are not artifacts of cDNA library construction. The blots also showed the presence of polydisperse transcripts in bovine submaxillary gland but no detectable hybridization signals in liver or brain RNA.**

Mucins are heavily glycosylated proteins with saccharide chains distinguishable in both composition and mode of linkage to the core protein. Mucins are primarily responsible for the gel character and biological functions of the viscous, sticky, gel-like mucus secretions that are produced by various internal cavities (oral and uterine) and tracts (respiratory and gastrointestinal) of higher vertebrates (1, 2). Mucus secretions serve common basic functions as lubricant and protective layer for the underlying epithelium that are in contact with the external environment. Even though the physico-chemical properties and functions of mucous secretions are attributed to the mucins, presently our understanding of the true polymeric structure (macrostructure) of these molecules is incomplete. Several studies suggest that mucin consists of subunits that are held together by covalent (disulfide) and/or noncovalent interactions; however, there is no agreement regarding the models proposed.

We are interested in elucidating the primary "subunit" structure of bovine submaxillary mucin (BSM). In our previous studies we identified a serine- and threonine-rich 60-kDa precursor protein that was immunoprecipitated by anti-apo-BSM from the *in vitro* translation products of bovine submaxillary gland mRNA (3). In the present study, a 2.0-kilobase (kb) cDNA clone encoding for a 58.9-kDa protein was isolated[§] from a λgt11 cDNA library prepared from bovine submaxillary gland mRNA. The deduced sequence of the cloned protein revealed two distinct domains. The 224-amino acid sequence from the carboxyl terminus was rich in cysteine, whereas the amino-terminal segment of 339 residues had a mucin-like composition. RNA-blot analysis indicates that submaxillary glands, but not liver or brain, contain the mRNA for this protein.

The expression of a cysteine-rich precursor protein in submaxillary gland is of considerable interest because previously structures of bovine (and ovine) submaxillary mucins were believed not to involve disulfide bonds (4, 5). In a preliminary abstract we have previously reported the isolation and partial characterization of the λBSM7 clone (6).

METHODS

Materials. Anti-apo-BSM antibodies were prepared and characterized in previous studies (3). Expression vector λgt11 was provided by Richard Hynes (Massachusetts Institute of Technology, Boston). *Escherichia coli* host strains Y1089R$^-$ and Y1090R$^-$ were purchased from Promega. Actin DNA was prepared from an actin cDNA plasmid provided by Melvin Billingsley of this institution.

Preparation of mRNA. Bovine submaxillary glands, liver, and brain removed immediately after sacrifice of the animal were brought in dry ice to the laboratory, cleaned free of blood by rinsing with cold phosphate-buffered saline, and dissected to remove fat and connective tissue. One gram of the cleaned tissues was used within 60–90 min of removal from the animal for the isolation of total RNA by using the guanidinium thiocyanate procedure as described by Han *et al.* (7). Poly(A)$^+$ RNA was prepared from total RNA by chromatography on oligo(dT)-cellulose (8).

Construction of the cDNA Library. Synthesis of the first cDNA strand was done by using the poly(A)$^+$ RNA (2 μg) as template and oligo(dT)$_{12-18}$ as primer for reverse transcription, and the second cDNA strand was synthesized by using RNase H and DNA polymerase I (9). The cDNA was methylated with *Eco*RI methylase; the *Eco*RI linkers (pGGAATTCC) were ligated, cut with *Eco*RI endonuclease, separated by gel filtration, and ligated into the unique *Eco*RI restriction site of λgt11 (10). The recombinant cDNA mole-

Abbreviations: BSM, bovine submaxillary mucin; IPTG, isopropyl β-D-thiogalactoside.

*Present address: Department of Human Genetics, Yale University School of Medicine, New Haven, CT 06510.
†Present address: Department of Biochemistry and Molecular Biology, Georgetown University School of Medicine, Washington, DC 20007.
‡To whom reprint requests should be addressed.

§The sequence reported in this paper has been deposited in the GenBank data base (accession no. M36192).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

cules were packaged into phage by using an *in vitro* packaging kit supplied by Amersham.

Immunological Screening. Antibodies from the rabbit antiserum (3) were affinity-purified on a column of apo-BSM coupled to CNBr-activated (11) Sepharose. These affinity-purified antibodies were preincubated with *E. coli* Y1089R$^-$(λgt11) lysogen extract, and the absorbed antibodies were used to screen the recombinants by the plaque method (12).

Subcloning and DNA Sequencing. The 1.8-kb and the 2.0-kb cDNA inserts of the *Eco*RI restriction endonuclease-digested recombinant phage DNA were separated by agarose gel electrophoresis and eluted by the freeze–squeeze procedure (13). They were subcloned into M13mp19 or M13mp18 bacteriophages (14) and sequenced by the dideoxynucleotide chain-termination procedure by using 17-mer Universal primer or 15-mer synthetic primers and deoxyadenosine 5′-[α-[^{35}S]thio]triphosphate as the label (15).

The cDNA inserts were further fragmented with the restriction endonucleases *Rsa* I and *Alu* I, and the fragments were subcloned in the *Sma* I restriction site of either M13mp18 or M13mp19 phage and sequenced. Sequences not obtained by subcloning of the restriction fragments were obtained by using chemically synthesized oligonucleotide primers. The DNA sequences were constructed and analyzed by using DNA and protein sequence analysis software package from IntelliGenetics.

Immunoblotting of the Lysates of Induced Recombinant Lysogens. *E. coli* Y1089R$^-$ cells were lysogenized either with the vector or the recombinant phage by infecting them at a multiplicity of infection of 1 and selecting the lysogens by their temperature sensitivity (10). Lysogens were grown at 32°C to a cell density of 2×10^8, incubated at 42°C for 15 min with the addition of 10 mM isopropyl β-D-thiogalactoside (IPTG) and again grown at 38°C for an additional 2 hr. Cells were pelleted, solubilized in $NaDodSO_4$-containing buffer, and subjected to gel electrophoresis.

Primer Extension. A 30-base oligonucleotide of sequence 5′-ACTCCCAAGTTGTTCTCCCGGGATC CTGTG-3′ was used to prime bovine submaxillary gland mRNA near its 5′ end. The synthetic oligonucleotide was purified by thin-layer chromatography (16) and end-labeled. The labeled primer was hybridized to various amounts of RNA and extended by reverse transcriptase under conditions of cDNA synthesis (17). The reaction mixture was treated with ribonuclease and phenol/chloroform-extracted before ethanol precipitation of the extension products.

Synthesis of mRNA from cDNA. The 1.8-kb and the 2.0-kb cDNA *Eco*RI fragments were cloned at the *Eco*RI site of the vector pBluescript II KS(−) (Stratagene). The 1.8-kb insert-carrying plasmid was cut with *Bam*HI or *Hin*dIII and transcribed using the RNA polymerases T3 and T7, respectively. The 2.0-kb insert-carrying plasmid was linearized with *Hin*dIII or *Not* I and transcribed by using RNA polymerase T7 and T3, respectively. The integrity of the RNA synthesized was checked by agarose gel electrophoresis and quantitated by incorporation of [α-^{32}P]UTP in a parallel reaction.

Cell-Free Translation. Translation of the synthetic mRNA by using rabbit reticulocyte lysate (Promega Biotec) as well as immunoprecipitation and gel electrophoretic analysis of the products were done as described (3).

RNA Blot Analysis. Total RNA (10 μg) and poly(A)$^+$ RNA (2 μg) preparations were subjected to electrophoresis on 1% agarose/formaldehyde denaturing gels, transferred to nitrocellulose membrane, and hybridized with DNA probes under stringent conditions, as described (18). The final wash of the hybridized filters was at 65°C with 15 mM NaCl/1.5 mM sodium citrate, pH 7.0/0.1% $NaDodSO_4$. The 1.8-kb cDNA from λBSM7 clone and the 877-base-pair (bp) and 729-bp fragments isolated by treatment of the 1.8-kb DNA with *Ava* II and actin cDNA were labeled by random-priming and used as hybridization probes for the RNA blots.

RESULTS

Construction and Screening of the Submaxillary Gland cDNA Library. Upon screening the cDNA library for the expression of the fusion protein immunoreactive with the affinity-purified anti-apo-mucin antibodies, seven clones were obtained, one of which was plaque-purified by rescreening and designated as λBSM7. This clone had a cDNA-insert fragment that was sequenced and found to be 1768 bp (≈1.8 kb) long. The nucleotide sequence included an open reading frame starting from the first nucleotide and, therefore, this clone was considered incomplete. On rescreening the library by using the 1.8-kb DNA as probe, three positive clones were obtained, one of which (λBSM10) had an insert of 2.0 kb.

Analysis of the Fusion Protein Produced by λBSM7. The IPTG-induced λgt11 lysogen produced the β-galactosidase protein, and the λBSM7 lysogen produced a distinct protein of 171 kDa (data not shown). The proteins on the gel were electroblotted to nitrocellulose filter and probed with anti-apo-BSM antiserum preabsorbed with the *E. coli* Y1089R$^-$ cell lysate. Alkaline phosphatase-conjugated goat anti-rabbit IgG antibody was used to visualize the immunoreactive proteins. Only the 171-kDa protein produced by the IPTG-induced lysogen of λBSM7 reacted with the anti-apo-BSM antibody, thus confirming the identity of the clone.

Nucleotide Sequence and Predicted Amino Acid Sequence of the Encoded Protein. The complete cDNA sequence and the encoded amino acid sequence of the λBSM10 clone is shown in Fig. 1. The cDNA sequence is 1972 bp long and includes the 5′-untranslated region, an open reading frame, the 3′-noncoding region followed by a poly(A) stretch of 20 nucleotides, and the flanking *Eco*RI linkers. The hexanucleotide AATAAA preceding the poly(A) tail by 16 nucleotides is probably the polyadenylylation signal because in most of the eukaryotic messages this signal precedes the poly(A) tract by 11–26 nucleotides (19). The A+T content of the coding region is 55%, and that of the noncoding region excluding the poly(A) stretch is 70%.

The open reading frame present in the cDNA sequence has the first ATG codon starting at the 72nd nucleotide, which codes for a protein with 563-amino acid residues. The calculated molecular mass of 58,920 Da for this protein is in good agreement with the molecular mass of 60 kDa estimated for the protein immunoprecipitated in our cell-free translation studies (3).

The first five bases upstream of the initiating ATG codon are not in agreement with the consensus CCRCC sequence (where R represents adenine or guanine), particularly the presence of a purine nucleoside at position 3, proposed by Kozak (20). However, an in-frame termination codon appears upstream at residues 15–17, suggesting that the ATG starting at nucleotide 72 is the initiator codon and that the BSM mRNA probably belongs to the minor (≈3%) class of mRNAs having a pyrimidine nucleoside at position −3 to the initiating AUG. To confirm the initiator site and to establish that the cDNA is full-length, we carried out primer-extension experiments and cell-free translation of mRNA synthesized from the 1.8-kb and 2.0-kb cDNAs in the sense and antisense orientations. Primer extension of bovine submaxillary gland mRNA with an oligonucleotide complementary to positions 21–50 in Fig. 1 showed extension products in the size range of 56–58 bp, which were incompletely resolved (Fig. 2); this represents 26- to 28-bp extensions of the primer. Gel electrophoresis of the translation mixtures of synthetic mRNAs showed a protein of ≈58 kDa only with the mRNA prepared from the 2.0-kb cDNA in the sense orientation. It should be noted that the mRNA obtained from the 1.8-kb cDNA in the

```
                GAATT CCC AAC TGG TAA AAT CAC AGG ATC CCG GGA GAA CAA CTT GGG AGT CAG GAA GTG AAG TTG CCA CAT 71

Met Lys Val Leu Gln Glu Asn Ser Pro Arg His Ala Ile Ser Gly Ser Ser His Thr Glu Ala Thr Thr Leu Ile Val Ser 27
ATG AAG GTA CTT CAG GAA AAT TCT CCA AGG CAC GCC ATA TCT GGG AGT TCC CAC ACA GAG GCC ACA ACT TTA ATA GTG AGC 152
 *
Asn Ser Thr Ser Gly Thr Gly Leu Arg Pro Glu Asp Asn Thr Ala Val Ala Gly Gly Gln Ala Thr Gly Arg Val Thr Gly 54
AAC AGC ACA TCT GGA ACT GGG CTC AGA CCT GAA GAT AAC ACT GCA GTA GCA GGA GGC CAA GCG ACT GGG CGT GTG ACA GGT 233

Thr Thr Lys Val Ile Pro Gly Thr Thr Val Ala Pro Gly Ser Ser Asn Thr Glu Ser Thr Thr Ser Leu Gly Glu Ser Arg 81
ACC ACA AAG GTA ATC CCT GGC ACA ACT GTT GCC CCT GGC AGT TCC AAC ACA GAG TCC ACA ACT TCC TTA GGA GAA AGT AGA 314

Thr Arg Ile Gly Arg Ile Thr Gly Ala Thr Thr Gly Thr Ser Lys Arg Ser Ser Pro Gly Ser Lys Thr Gly Asn Thr Gly 108
ACA AGA ATT GGA AGA ATC ACA GGA GCC ACC ACT GGC ACA TCT AAG AGG TCA AGC CCT GGA AGT AAA ACA GGT AAC ACT GGT 395

Ala Ile Ser Gly Thr Thr Val Ala Pro Gly Ser Ser Asn Thr Gly Ala Thr Thr Ser Leu Gly Ser Gly Glu Thr Thr Gln 135
GCA ATC TCT GGC ACA ACA GTT GCA CCC GGA AGC TCT AAC ACA GGG GCT ACA ACT TCT TTG GGA AGT GGT GAA ACC ACC CAG 476

Gly Gly Ile Lys Ile Val Thr Met Gly Val Thr Thr Gly Thr Thr Ile Ala Pro Gly Ser Ser Asn Thr Lys Ala Thr Thr 162
GGT GGA ATT AAA ATA GTT ACC ATG GGG GTG ACT ACT GGT ACA ACC ATT GCA CCT GGA AGT TCG AAC ACA AAG GCT ACA ACT 557

Pro Thr Glu Val Arg Thr Thr Thr Glu Val Arg Thr Ala Thr Glu Thr Thr Thr Ser Arg His Ser Ser Asp Ala Thr Gly 189
CCT ACA GAA GTC AGA ACT ACC ACT GAA GTC AGA ACA GCA ACT GAA ACT ACC ACT TCA AGA CAC AGT AGT GAT GCC ACT GGG 638

Ser Gly Ile Gln Thr Gly Ile Thr Gly Thr Gly Ser Gly Thr Thr Ser Ser Pro Gly Gly Phe Asn Ala Glu Ala Thr Thr 216
AGT GGA ATA CAA ACA GGT ATC ACT GGG ACA GGC TCT GGA ACT ACA TCC TCA CCT GGA GGT TTC AAT GCA GAA GCA ACA ACT 719

Phe Lys Glu His Val Arg Thr Thr Glu Thr Arg Ile Leu Ser Gly Thr Thr Arg Gly Arg Ser Gly Thr Thr Val Ile Pro 243
TTT AAA GAA CAT GTT AGA ACC ACT GAA ACA AGA ATT TTA TCA GGT ACC ACT AGG GGA CGC TCT GGC ACA ACA GTT ATT CCT 800

Glu Ser Ser Asn Thr Gly Thr Ser Thr Gly Val Gly Arg Gln Thr Ser Thr Ala Val Val Ser Gly Arg Val Thr Gly Val 270
GAA AGT TCC AAC ACA GGG ACT AGC ACT GGA GTT GGA AGG CAA ACA AGT ACT GCT GTG GTA TCA GGC AGA GTT ACT GGT GTC 881

Ser Glu Ser Ser Ser Pro Gly Thr Ser Lys Glu Ala Ser Glu Thr Thr Thr Gly Pro Gly Ile Ser Thr Thr Gly Ser Thr 297
TCA GAA AGT TCC AGC CCA GGT ACC TCC AAG GAA GCA TCT GAA ACA ACT ACT GGT CCT GGG ATT TCT ACC ACT GGC TCC ACT 962

Ser Lys Ser Asn Arg Ile Thr Thr Ser Ser Arg Ile Pro Tyr Pro Glu Thr Thr Val Val Ala Thr Gly Glu Gln Glu Thr 324
TCA AAA TCA AAT AGA ATC ACA ACA AGC TCC AGA ATA CCC TAC CCA GAA ACC ACT GTT GTA GCA ACA GGA GAA CAA GAA ACT 1043

Glu Thr Lys Thr Gly Cys Thr Thr Ser Leu Pro Pro Pro Pro Ala Cys Tyr Gly Pro Leu Gly Glu Lys Lys Ser Pro Gly 351
GAA ACT AAG ACA GGA TGT ACA ACA TCT CTT CCA CCA CCT CCA GCT TGT TAT GGT CCA CTG GGA GAA AAG AAG TCA CCT GGA 1124

Asp Ile Trp Thr Ala Asn Cys His Lys Cys Thr Cys Thr Asp Ala Glu Thr Val Asp Cys Lys Leu Lys Glu Cys Pro Ser 378
GAC ATA TGG ACT GCC AAT TGC CAC AAA TGC ACC TGT ACT GAT GCA GAG ACT GTA GAC TGT AAA CTC AAG GAG TGT CCT TCT 1205
                                                  *
Pro Pro Thr Cys Lys Pro Glu Glu Arg Leu Val Lys Phe Lys Asp Asn Asp Thr Cys Cys Glu Ile Ala Tyr Cys Glu Pro 405
CCA CCC ACA TGC AAA CCT GAA GAG AGA CTT GTA AAG TTC AAA GAT AAT GAT ACC TGC TGT GAA ATT GCA TAC TGT GAA CCA 1286

Arg Thr Cys Leu Phe Asn Asn Asn Asp Tyr Glu Val Gly Ala Ser Phe Ala Asp Pro Lys Asn Pro Cys Ile Ser Tyr Ser 432
AGA ACA TGT TTA TTT AAC AAT AAT GAC TAT GAG GTT GGT GCT TCA TTT GCT GAC CCT AAG AAC CCG TGT ATC TCT TAC TCC 1367

Cys His Asn Thr Gly Phe Val Ala Val Val Gln Asp Cys Pro Lys Gln Thr Trp Cys Ala Glu Glu Asp Arg Val Tyr Asp 459
TGC CAC AAC ACT GGT TTC GTT GCC GTG GTT CAA GAC TGC CCG AAG CAG ACC TGG TGT GCA GAA GAA GAC AGA GTC TAT GAT 1448
                                                        *
Ser Thr Lys Cys Cys Tyr Thr Cys Lys Pro Tyr Cys Arg Ser Ser Ser Val Asn Val Thr Val Asn Tyr Asn Gly Cys Lys 486
TCA ACA AAA TGT TGC TAT ACA TGT AAA CCT TAT TGC AGA TCT TCA TCC GTG AAT GTG ACT GTT AAC TAT AAT GGT TGC AAG 1529

Lys Lys Val Glu Met Ala Arg Cys Ala Gly Glu Cys Lys Lys Thr Ile Lys Tyr Asp Tyr Asp Ile Phe Gln Leu Lys Asn 513
AAA AAA GTT GAG ATG GCA AGA TGC GCA GGG GAA TGC AAG AAA ACC ATC AAG TAT GAT TAT GAC ATC TTT CAG TTG AAA AAT 1610

Ser Cys Leu Cys Cys Gln Glu Glu Asn Tyr Glu Tyr Arg Glu Ile Asp Leu Asp Cys Pro Asp Gly Gly Thr Ile Pro Tyr 540
TCA TGC CTT TGC TGC CAA GAA GAA AAC TAT GAG TAT AGG GAA ATT GAT CTT GAC TGT CCT GAT GGT GGT ACA ATA CCA TAT 1691

Arg Tyr Arg His Ile Ile Thr Cys Ser Cys Leu Asp Ile Cys Gln Gln Ser Met Thr Ser Thr Val Ser TER           563
AGG TAC AGG CAT ATC ATT ACA TGT TCC TGT TTA GAC ATA TGC CAA CAG TCT ATG ACT TCA ACA GTC AGT TAA TAA TAA ATG 1772

GCA TTA CCT TTC CTG CTG TAA CAG ACC AGG TAT TTA TTT TAT AAG TGA ACA AAA ATA ATG ACT AAA TTG TGA AAA TAA CTA 1853

AAC ATT TAC ATA TTG CTC AAA AAG TGA GTT GTC ATT ATG AGG TTT TTT GTT CTT TTA TTG ACT GGA TCT AAA ATA TAA ATA 1934

CAA CTT TTC AAG CAA AAA AAA AAA AAA AAA AAA GGA AT                                                    1972
```

FIG. 1. Nucleotide sequence and the deduced amino acid sequence of the λBSM10 cDNA clone. The 11-amino acid-homologous sequence present thrice and the pentapeptide following two of these are underlined. The hexanucleotide sequences AATAAA preceding the polyadenylylation site are boxed. The potential N-glycosylation sites are marked with asterisks. The proline tetrapeptide and lysine dipeptide are overlined. The arrow marks the arbitrary position, before the second cysteine residue, at which the protein is separated into the mucin-like and cysteine-rich domains.

sense orientation did not yield a translation product, confirming that this cDNA is incomplete.

RNA Blot Analysis. Fig. 3*A* shows the result of blot hybridization analysis of total RNA from bovine liver, brain, and submaxillary gland by using the 1.8-kb DNA as well as the *Ava* II fragments derived from it as probes. All three probes hybridized to a heterogenous population of RNA, indicating multiple transcripts. Even after a 32-fold-longer exposure of the blot to x-ray film, no hybridization signal was detected to liver and brain RNA. Poly(A)$^+$ RNA from bovine submaxillary gland also showed a family of transcripts with the three probes (Fig. 3*B*). In the mRNA blot hybridized with the 1.8-kb DNA probe, the presence of at least six distinct signals ranging in size from 1.0 to 2.9 kb is evident (Fig. 3*B*). Both total and poly(A)$^+$ RNA gave one strong signal when probed with actin DNA probe, indicating that the polydisperse nature of the signals noted with the other probes is not from degradation of the RNA preparations.

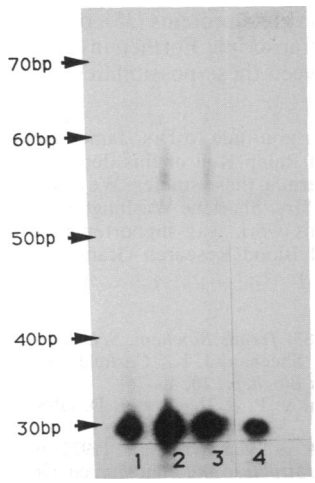

FIG. 2. Primer-extension products of bovine submaxillary total RNA. The reactions were done as described in the text with 150 μg, 20 μg, and 10 μg of total RNA (lanes 1, 2, and 3, respectively), and the products were analyzed on a 6% polyacrylamide gel. Lane 4 is the minus RNA control showing the unextended 30-bp primer. The major extension products are in the 56- to 58-bp size range.

Homology Search of Amino Acid Sequence. Searches of the deduced submaxillary gland apoprotein sequence with other protein sequences registered in the National Biomedical Research Foundation Protein Identification Resource (PIR; release 23) and the University of Geneva Protein (Swiss-Prot; release 13) sequence data banks revealed statistically significant matches of 10 SDs or more to yeast glucoamylase S1 fragment (21) and precursor (22), pig submaxillary gland apomucin (23), and mouse period clock protein (24). Of these matches, the one of most interest is the match between the mucin-like amino-terminal segment 1–282 of the cloned bovine submaxillary apoprotein and the porcine apomucin fragment (225–503) (23). There is an exact match of 104 residues (37%), and the similarity between the sequences increases to 164 residues (58%) when conservative amino acid substitutions were also taken into account.

DISCUSSION

In this study, we have cloned and sequenced a 2.0-kb cDNA encoding a bovine submaxillary protein of deduced molecular mass of 58,920 Da. Even though the nucleotide sequence around the ATG at positions 72–74 of this cDNA does not match Kozak's consensus sequence, evidence from primer extension and translation of RNA synthesized from the 2.0-kb cDNA strongly suggests that the above ATG is the initiator codon and that the 2.0-kb cDNA is full-length. The protein product of this cDNA cross-reacts with polyclonal antibody raised against the deglycosylated major component of the bovine submaxillary mucin, but its deduced amino acid sequence reveals a region of very high cysteine content, a feature unexpected for a mucin core protein (25, 26). The results of the RNA blot analysis suggest (*i*) that this protein is abundant in submaxillary gland and is specific to this tissue because no detectable hybridization signal occurred in liver or brain. (*ii*) That the 729-bp DNA probe derived from the mucin-like region of the protein and the 877-bp DNA probe derived from the cysteine-rich region of the protein gave identical hybridization signals excludes the possibility that the clone was an artifact from splicing of two different cDNA fragments during reverse transcriptase reading of mRNA.

A striking feature of the deduced protein sequence is the skewed distribution of the amino acids, most notably that of the hydroxy amino acids and cysteine. There are 30 cysteines in the molecule but none in the first 329 residues. The first cysteine at position 330 is followed by three hydroxyamino acids, leucine, a unique proline tetrapeptide, and alanine before the next cysteine. Because prolines are known to be helix breakers, the presence of a proline tetrapeptide in the region demarcating the mucin-like and cysteine-rich domains may be especially significant in terms of the secondary structure of this protein. There is also a pair of lysines at positions 347 and 348, which may be a proteolytic processing signal analogous to those found in mammalian hormone precursors and yeast glucoamylase, as pointed out by Yamashita *et al.* (22). On arbitrarily dividing the protein between alanine at position 339 and cysteine at position 340, two polypeptide fragments of contrasting composition are obtained (Table 1). The amino-terminal fragment of 339 residues is very rich in threonine, serine, and glycine (these three constituting ≈53.7 mol % of amino acids) and poor in cysteine, aspartic acid, tyrosine, phenylalanine, and tryptophan (these five constituting only 1.8 mol %). Most threonines and serines (92 of 134) are present in clusters of two, three, or four, and a further 24 are followed by glycine. The tendency for serine and threonine residues to be adjacent to themselves or to one another and to

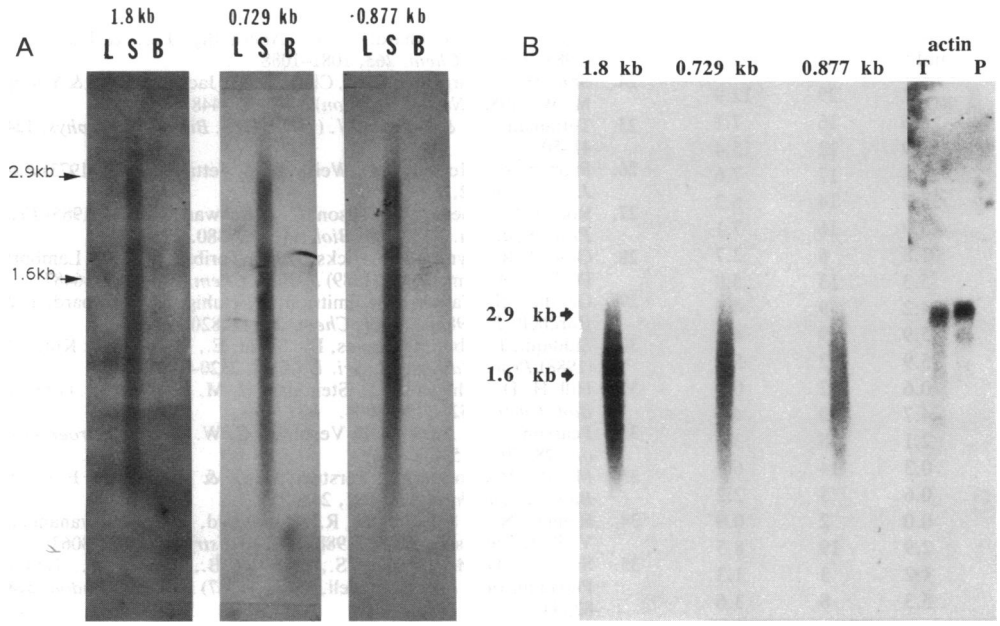

FIG. 3. RNA-blot analysis as described in the text. (*A*) Total RNA isolated from bovine liver (L), submaxillary gland (S), and brain (B) hybridized with the 1.8-kb cDNA isolated from λBSM7 and the two fragments (0.729 kb and 0.877 kb) isolated by *Ava* II treatment of the 1.8-kb DNA. (*B*) Poly(A)$^+$ RNA isolated from bovine submaxillary gland hybridized with the same three DNA probes (lanes 1–3). Total (T) and poly(A)$^+$ RNA (P) hybridized with actin cDNA (right lanes). Positions of RNA molecular length markers are indicated.

be near glycine has also been noted in the deduced partial sequence for porcine submaxillary mucin (23) and the proteoglycan core proteins (27).

Of the ten asparagine residues in the first 339 residues, eight are followed by threonine or serine and of these eight one is present in the consensus N-glycosylation sequence (Asn-Xaa-Ser/Thr) (Fig. 1). In the mucin-like segment of the molecule, an 11-amino acid sequence is repeated thrice with valine being replaced by isoleucine in the third repeat. The first two repeats are also followed by a homologous pentapeptide Thr-Thr-Ser-Leu-Gly (Fig. 1). There is no homology in the nucleotide sequence of these three repeats. Recent investigations seem to indicate the presence of tandem repetitive sequences in mucin core proteins. Thus, repetitive sequences of 81, 23, and 20 amino acids have been noted in the core proteins of porcine submaxillary mucin (23), human intestinal mucin (28), and breast cancer-associated mucin glycoproteins (29, 30), respectively. But interestingly, the partial amino acid sequence of the ovine submaxillary mucin failed to reveal any repetitive sequences (31).

In contrast to the amino-terminal segment, the carboxyl-terminal 224-amino acid segment of the cloned protein was rich in cysteine, aspartic acid, tyrosine, and lysine (these four constitute 34.7 mol %) and comparatively poorer in threonine, serine, and glycine (total of 17.8 mol %). The uniqueness of this high concentration of cysteine in any protein is also indicated by our failure to detect statistically significant (within 8 SDs) homology between the 224-amino acid carboxyl-terminal segment of the cloned protein and the protein sequences registered in the searched data banks.

The cloned protein could be the putative link protein of BSM analogous to the link proteins proposed for pig gastric mucin (32), human small intestinal mucin (33), and tracheobronchial mucins of human and dog (34). Purification and characterization of the cloned protein from bovine submaxillary gland would provide more definitive information on the role of this protein in the BSM structure. Finally, the RNA-blot analysis under hybridization conditions of high stringency shows polydispersity of the submaxillary gland transcripts for this protein. Similar polydispersity was also observed for human intestinal mucin message (28). The polydispersity could be due to the presence of a multigene family, as in breast cancer-associated mucin glycoproteins (35), or differential processing of the mRNA transcript. Further investigation is needed to distinguish between these possibilities.

Table 1. Amino acid composition of the cloned bovine submaxillary gland apoprotein and its two domains

Amino acid	Apoprotein		Mucin-like domain		Cysteine-rich domain	
	n	mol %	*n*	mol %	*n*	mol %
Cys	30	5.3	1	0.3	29	12.9
Asp	18	3.2	2	0.6	16	7.1
Asn	22	3.9	10	2.9	12	5.4
Thr	101	17.9	84	24.8	17	7.6
Ser	64	11.4	50	14.7	14	6.2
Glu	36	6.4	20	5.9	16	7.1
Gln	12	2.1	6	1.8	6	2.7
Pro	31	5.5	18	5.3	13	5.8
Gly	57	10.1	48	14.2	9	4.0
Ala	29	5.2	20	5.9	9	4.0
Val	32	5.7	20	5.9	12	5.4
Met	4	0.7	2	0.6	2	0.9
Ile	26	4.6	16	4.7	10	4.5
Leu	15	2.7	7	2.1	8	3.6
Tyr	15	2.7	1	0.3	14	6.2
Phe	7	1.2	2	0.6	5	2.2
Trp	2	0.4	0	0.0	2	0.9
Lys	29	5.2	10	2.9	19	8.5
His	7	1.2	4	1.2	3	1.3
Arg	26	4.6	18	5.3	8	3.6

We express our gratitude to Drs. James Hopper, Anita Hopper, Charles Hill, and Ralph Keil of this department for many helpful discussions concerning these studies. We acknowledge the excellent technical help of Mrs. Sharlene Washington in performing the RNA blot analysis. This work was supported by National Institute of Heart, Lung and Blood Research Grants HL42651 (V.P.B.) and HL42332 (E.A.D.).

1. Allen, A. (1983) *Trends Biochem. Sci.* **8,** 169–173.
2. Carlstedt, I., Sheehan, J. K., Corfield, A. P. & Gallagher, J. T. (1985) *Essays Biochem.* **20,** 40–76.
3. Bhavanandan, V. P. & Hegarty, J. D. (1987) *J. Biol. Chem.* **262,** 5913–5917.
4. Gottschalk, A. & Bhargava, A. S. (1972) in *Glycoproteins: Their Composition, Structure and Function*, ed. Gottschalk, A. (Elsevier, Amsterdam), Vol. 5, part B, pp. 810–829.
5. Hill, H. D., Reynolds, J. A. & Hill, R. L. (1977) *J. Biol. Chem.* **252,** 3791–3798.
6. Bhargava, A. K., Bhavanandan, V. P. & Davidson, E. A. (1988) *FASEB J.* **2,** A1033 (abstr.).
7. Han, J. H., Stratowa, C. & Rutter, W. J. (1987) *Biochemistry* **26,** 1617–1625.
8. Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1408–1412.
9. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25,** 263–269.
10. Young, R. A. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1194–1198.
11. Shapiro, D. J., Taylor, J. M., McKnight, G. S., Palacios, R., Gonzalez, C., Kiely, M. L. & Schimke, R. T. (1974) *J. Biol. Chem.* **249,** 3665–3671.
12. Young, R. A. & Davis, R. W. (1983) *Science* **222,** 778–782.
13. Thuring, R. W. J., Sanders, J. P. M. & Borst, P. (1975) *Anal. Biochem.* **66,** 213–220.
14. Messing, J., Carlson, J., Hagen, G., Rubenstein, I. & Oleson, A. (1984) *DNA* **3,** 31–40.
15. Sanger, F., Nicklen, S. & Coulsen, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
16. Alvarado-Urbina, G., Sathe, G. M., Liu, W., Gillen, M. F., Duck, P. D., Bender, R. & Olgilvie, K. K. (1981) *Science* **214,** 270–274.
17. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) in *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), 2nd Ed., Vol. 1, pp. 7.79–7.87.
18. Davis, L. G., Dibner, M. D. & Battey, J. F. (1986) *Basic Methods in Molecular Biology* (Elsevier, New York), pp. 143–146.
19. Tosi, M., Young, R. A., Hagenbuchle, O. & Schibler, U. (1981) *Nucleic Acids Res.* **9,** 2313–2323.
20. Kozak, M. (1984) *Nucleic Acids Res.* **12,** 857–872.
21. Yamashita, I., Nakamura, M. & Fukui, S. (1987) *J. Bacteriol.* **169,** 2142–2149.
22. Yamashita, I., Suzuki, K. & Fukui, S. (1985) *J. Bacteriol.* **161,** 567–573.
23. Timpte, C. S., Eckhardt, A. E., Abernathy, J. L. & Hill, R. L. (1988) *J. Biol. Chem.* **263,** 1081–1088.
24. Shin, H. S., Bargiello, T. A., Clark, B. P., Jackson, F. R. & Young, M. W. (1985) *Nature (London)* **317,** 445–448.
25. Tettamanti, G. & Pigman, W. (1968) *Arch. Biochem. Biophys.* **124,** 41–50.
26. Pigman, W., Moschera, J., Weiss, M. & Tettamanti, G. (1973) *Eur. J. Biochem.* **32,** 148–154.
27. Roden, L., Koerner, T., Olson, C. & Schwartz, N. B. (1985) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **44,** 373–380.
28. Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lambort, D. T. A. & Kim, Y. S. (1989) *J. Biol. Chem.* **264,** 6480–6487.
29. Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J. & Burchell, J. (1988) *J. Biol. Chem.* **263,** 12820–12823.
30. Siddiqui, J., Abe, M., Hayes, D., Shani, E., Yunis, E. & Kufe, D. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2320–2323.
31. Hill, H. D., Schwyzer, M., Steinman, H. M. & Hill, R. L. (1977) *J. Biol. Chem.* **252,** 3799–3804.
32. Pearson, J. P., Allen, A. & Venables, C. W. (1980) *Gastroenterology* **78,** 709–715.
33. Mantle, M., Potier, M., Forstner, G. G. & Forstner, J. F. (1986) *Biochim. Biophys. Acta* **881,** 248–257.
34. Ringler, N. J., Selvakumar, R., Woodward, H. D., Bhavanandan, V. P. & Davidson, E. A. (1988) *Biochemistry* **27,** 8056–8063.
35. Swallow, D. M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J. & Bramwell, M. E. (1987) *Nature (London)* **328,** 82–84.