The Complete cDNA Sequence and Structural Polymorphism of the Polypeptide Chain of Porcine Submaxillary Mucin*

(Received for publication, July 24, 1997, and in revised form, October 11, 1997)

Allen E. Eckhardt‡, Candace S. Timpte§, Alexander W. DeLuca, and Robert L. Hill¶
From the Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710

The complete structure of the DNA encoding the polypeptide chain of porcine submaxillary mucin has been determined. The polypeptide is composed of distinct domains. A large central domain containing tandem repeats of 81 residues each is flanked by much shorter domains with sequences similar to the tandem repeats. Four disulfide-rich domains, three at the amino terminus and one at the carboxyl terminus, complete the chain. The disulfide-rich domains have significant sequence identity to those of other mucins and prepro-von Willebrand factor. The coding region of the mucin gene is highly polymorphic, and three alleles were identified in a single animal that encoded different numbers of the 81-residue tandem repeats. A single large exon devoid of introns encodes the tandem repeat domains. The largest allele with 135 tandem repeats encoded 13,288 amino acids to give a polypeptide with $M_r = 1,184,106$. The other two alleles contained 99 and 125 tandem repeats, respectively. Each allele also showed different restriction fragment length polymorphisms, which is consistent with the different patterns seen in individual animals. Fragment length polymorphism was also seen within two different families of animals, indicating that the polymorphism observed occurs in a single generation.

It is now established that different cells of the mammalian respiratory, gastrointestinal, and urogenital tracts secrete structurally unique mucins that differ not only in the amino acid sequence of their polypeptide chains but also in the structure of the oligosaccharides covalently bound to the polypeptide chains. Nevertheless, the complete amino acid sequence of only one large, secreted mammalian mucin (1) and two smaller, less complex mucins (2, 3) has been established, although portions of the sequences of many mucins have been determined recently (see, e.g., Refs. 4–7). From this structural information, it has also become clear that there are structural similarities among the mucins. All contain tandemly repeated sequences, although the number of amino acids in each tandem repeat and the number of repeats in each polypeptide chain is different. The amino acid sequence of a repeat bears no identity from one mucin to another, except that each repeat is rich in serine and threonine. It is thought that the vast majority of the oligosaccharides in mucins is in O-glycosidic linkage to the hydroxyl groups in the serine and threonine residues of the repeat domains. Moreover, many secreted mucins appear to contain disulfide-rich domains that flank the tandem repeat domains (1, 4–7). Many of these domains, called D-domains,† have significant amino acid sequence identity to one another and to those in human prepro-von Willebrand factor (1). The D-domains of von Willebrand factor are required for assembly of prepro-von Willebrand factor disulfide-bonded dimers into multimers, which interact with platelet receptors to initiate hemostasis through platelet adhesion at the site of vascular injury (8). Disulfide-bonded dimers of prepro-von Willebrand factor are formed through the carboxyl-terminal disulfide-rich domains that show little sequence similarity with D-domains.

The demonstration that the carboxyl-terminal disulfide-rich domain of porcine submaxillary mucin can form disulfide-bonded dimers supports the suggestion that the D-domains in mucins may also be essential for assembly of mucin multimers from disulfide-bonded dimers (11).

We report here the complete polynucleotide sequence that encodes the polypeptide chain of porcine submaxillary mucin. Portions of these structures have been described (12, 13), including the amino acid sequence of the carboxyl-terminal disulfide-rich domain, the amino acid sequence of five 81-residue tandem repeats and the sequence that joins the repeat domain to the carboxyl-terminal disulfide-rich domain. The new sequences described here account for about 90% of the mucin polypeptide, which is the largest polypeptide reported to date for a mammalian mucin.

EXPERIMENTAL PROCEDURES

Materials—Oligonucleotides were synthesized by the DNA Synthesis Facility at Duke University. The following materials were obtained commercially: DNA restriction endonucleases, T4 DNA ligase, T4 poly nucleotide kinase, Superscript II RNase H reverse transcriptase, ribonuclease H, terminal deoxynucleotidyltransferase, Taq DNA polymerase, 1-kb DNA ladder, high molecular weight DNA markers (8.3–48.5 kb), Glass-Max spin cartridges, ultra-pure agarose and cosmid mapping system (Life Technologies, Inc.); porcine genomic library in the pWE15 cosmid (average insert size 38 kb, CLONTECH); ScPlaque GTG low melting point agarose (FMC); TA cloning kit (Invitrogen Corp.); Long Range gel purification (AT Biochem); ultrapure deoxy nucleotides (Pharmacia Biotech Inc.); Sequenase Quick-Denature plasmid sequencing kit, [γ-32P]ATP (3000 Ci/mmol), and [35S]dATP (1000 Ci/mmol). Gel Filtration (35,000–40,000 M) (Sigma); ampicillin, sodium salt (Sigma); nitrocellulose filters (Schleicher & Schuell); Qiagen columns (mini-, midi-, maxi-, and manifold prep, Qiagen); DE-81 ion exchange paper (Whatman International Ltd.); Taq T7 reticulocyte lysate system (Promega Corp.); and [35S]lysine (1000 Ci/mmol, ICN).

Scanning of the Porcine Genomic Cosmid Library—The porcine genomic library in the pWE15 cosmid was screened by hybridization

* This work was supported in part by Research Grant 25766 (to R. L. H.) from the NIGMS, National Institutes of Health. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact. The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) AF005273.‡ Present address: Xanthom, Inc., Research Triangle Park, NC 27709.§ Present address: Dept. of Biological Sciences, University of New Orleans, New Orleans, LA 70148.¶ To whom correspondence should be addressed: Dept. of Biochemistry, P. O. Box 3711, Duke University Medical Center, Durham, NC 27710. Tel: 919-681-8805; Fax: 919-684-5040; E-mail: hill@biochem.duke.edu.

† The abbreviations used are: D-domain, disulfide-rich domain; kb, kilobase(s); nt, nucleotide(s); PCR, polymerase chain reaction.
for 50 min, followed by heating to 70 °C for 15 min to stop the reverse transcriptase for 10 min at 25 °C and then transferred to 45 °C. cDNA were hybridized and reverse transcribed (14) with Superscript II.

2.5 mM MgCl2, and 5 units of DNA polymerase. Thirty-five cycles of Taq DNA polymerase. Thirty-five cycles of Taq

RESULTS

The studies presented here give the complete nucleotide sequence of the DNA encoding porcine submaxillary mucin and the derived amino acid sequence of the mucin polypeptide chain. To facilitate the description of these studies, the structure of the polypeptide chain that was obtained is shown diagrammatically in Fig. 1, along with the domain structure of human intestinal mucin (MUC2) (1), which was the largest secreted, mammalian mucin previously described, and has many structural similarities to submaxillary mucin. The polypeptide chain of submaxillary gland mucin shown in Fig. 1 contains 13,288 amino acid residues and can be divided into different domains. There are three D-domains, which are formed by amino acids within residues 1–1278, and a structurally unrelated disulfi de-rich domain at the carboxyl terminus formed by residues 13059–13288. The tandem repeat domain extends from residue 1571 to residue 12529 and is the largest domain in the molecule. The tandem repeat domain varies somewhat in length, as discussed below, but the longest domain identified is composed of 135 81-residue repeats. The repeats are rich in glycine, serine, threonine, and alanine, which account for 75% of the amino acids in the domain. Flanking the tandem repeats are unique domains (residues 1279–1570 and residues 12530–13058) that have no sequence identity to one another or to any other domain in any mucin, but have an amino acid composition that is qualitatively much like that of the tandem repeat domain.

The complete amino acid sequence of the mucin polypeptide chain and the nucleotide sequence of the gene encoding it will not be given here, although portions of the sequence have been reported previously (12, 13). However, the complete nucleotide and amino acid sequences have been deposited in the GenBank™ data base, and are readily available by hyperlink in the template and two primers near the 5’ end of the deduced cDNA sequence. The primers, which were complementary to bases 85–103 and 73–94, were used for reverse transcription and PCR, respectively. The amplified PCR product was subcloned into the pCI II vector and the DNA inserts sequenced.

The complete genomic library was screened in succession with the following two primers, 5’-GGCTACAACTTCCATAGAAGG, which hybridized three nucleotides downstream from the last tandem repeat. Three colonies were selected which were positive for both probes (Ap050, Ap060, and Ap070) and were DNA prepared as above.

Mapping of Genomic DNA Inserts—Repeat positive cosmids were mapped for restriction sites by the procedure of Rackwits et al. (15). Briefly, cosmids were partially or completely digested with BamHI and the digests divided in half and hybridized to either oligonucleotide L (5’-AGGTGCGCCGCC) or oligonucleotide R (5’-GGCCGCGCACCT), which hybridize, respectively, to the left or right single-stranded cos termini. The size of a partially digested DNA fragment detected by hybridization with oligonucleotide L represented the distance from a restriction site to the left cos terminus. In this manner, the restriction sites for each of the cloned cosmid genomic DNA inserts could be mapped.

DNA Sequence Flanking the Repeat Region—The repeat positive cosmid clones from the first round of screening were digested with NotI and the full-length inserts purified by gel electrophoresis in low melting point agarose. All of the inserts were further digested with PstI, and the large resistant fragments that flanked the repeat regions were isolated by gel electrophoresis, cloned into pEMBL, and sequenced. Both strands were sequenced by methods used previously (12). None of the cosmids identified by screening with the repeat monomer contained the entire repeat region.

Anchored PCR to Obtain the 5’ End of Submaxillary Mucin cDNA—Total RNA was prepared from a porcine submaxillary gland as described previously (12). Total RNA (11 μg in 28 μl) and a primer (3.5 nmol) complementary to bases 4674–4702 of the submaxillary mucin cDNA were hybridized and reverse transcribed (14) with Superscript II reverse transcriptase for 10 min at 25 °C and then transferred to 45 °C for 50 min, followed by heating to 70 °C for 15 min to stop the reaction. After cooling on ice for 1 min, 1 μl of RNase H II (2 units) was added and the mixture incubated at 37 °C for 20 min. Single-stranded cDNA was purified from the reaction mixture using Glass-Max spin columns and the DNA eluted with 40 μl of 10 mM Tris-HCl (pH 8.0), 1 mM EDTA.

The single-stranded cDNA was G-tailed at the 3’ end with terminal deoxynucleotidyltransferase as follows. Four μl of single-stranded cDNA prepared as described above was incubated in 0.1 n potassium cacodylate (pH 7.2), 2 mM cobalt chloride, 0.2 mM dithiothreitol, 10 μM dGTP, and 15 units of terminal transferase in a final volume of 50 μl at 37 °C for 30 min. The poly-G-tailed single-stranded cDNA was purified using Glass-Max spin columns and eluted in 40 μl of 10 mM Tris-HCl (pH 8.0), 1 mM EDTA.

The poly-G-tailed single-stranded cDNA was amplified by PCR in an Ericom Thermocycler using an anchor primer complementary to the poly-G tail (5’-AAGCTTGGATCCCTAGAATTCCTCCCTCCCTCCCTCCC-3’) and a primer complementary to bases 4324–4362 of the submaxillary mucin cDNA. The PCR reaction mixture (100 μl) contained 5 μl of poly-G-tailed single-stranded cDNA, 20 pmol of each primer, 20 μM Tris-HCl (pH 8.4), 50 mM KCl, 0.4 mM dATP, dCTP, dGTP, and dTTP, 2.5 mM MgCl2, and 5 units of Taq DNA polymerase. Thirty-fi ve cycles of amplification were performed with a denaturation temperature of 94 °C for 2 min, an annealing temperature of 55 °C for 30 s, and an extension temperature of 72 °C for 3 min.

The amplified products were subcloned into the pCRII vector as recommended by the supplier (Invitrogen Corp.). Plasmids were prepared from positive colonies, and the inserts sized by agarose gel electrophoresis and sequenced as above. From the deduced sequence, a new set of primers was synthesized for another cycle of anchored PCR. A total of six cycles was necessary to obtain the entire DNA sequence for porcine submaxillary mucin. The primers used for each cycle of anchored PCR were complementary to the following submaxillary mucin cDNA sequences: cycle 2, bases 4255–4274, 4239–4258, and 4041–4058; cycle 3, bases 2927–2948, 2877–2900, and 2820–2839; cycle 4, bases 2208–2226, 2193–2215, and 1970–1987; cycle 5, bases 988–1012, 962–985, and 894–915; cycle 6, bases 638–660, 625–642, and 577–598.
Porcine Submaxillary Mucin Structure

FIG. 1. Schematic structure of the domains in porcine submaxillary mucin and human intestinal mucin (MUC2). The domains of porcine submaxillary mucin and human intestinal mucin are indicated as described previously (1). The lengths of the two polypeptides are to scale. The unique sequence domains of submaxillary mucin are unrelated to one another in sequence or to the sequence of the threonine/serine/proline-rich (TSP-Rich) domain of MUC2.

electronic version of the Journal.\(^2\) Previous studies gave the amino acid sequence of the carboxyl-terminal cystine-rich domain (13), several of the 81-residue tandem repeats (12), and the unique sequence domain that joins these two domains. Thus, further studies were undertaken to determine the sequence amino-terminal to the tandem repeats and the number of tandem repeats in the chain.

Determination of the Amino Acid Sequence Immediately Amino-terminal to the Tandem Repeat Domain—Because of the great length of the DNA encoding the submaxillary mucin, clones that encode the entire length of the DNA were never found in the Agt11 library prepared from porcine submaxillary mRNA (12). Thus, to extend the sequence amino-terminal to the tandem repeat domain, it was necessary to try to obtain this region, or some portion of it, from genomic DNA. For this purpose, a genomic DNA library in the cosmids pWE15 was screened with oligonucleotides encoding the tandem repeat sequence. Twenty cosmids were identified and purified, and their inserts, ranging in size from 33 to 43.5 kb, obtained after digestion with NotI. No inserts contained the entire repeat domain. The inserts contained varying numbers of tandem repeats and overlapped in sequence with nucleotides either 5' or 3' to the tandem repeats. The longest of these with sequence 5' to the tandem repeats, designated Apo1 (35.6 kb), was digested with PstI. There is one PstI site in each tandem repeat (12); thus, digestion of Apo1 with PstI gave the 243-nt DNA encoding the tandem repeats and a 3.6-kb DNA encoding the region 5' to the repeats. Sequence analysis of the 3.6-kb DNA showed it to encode 170 nt of the first tandem repeat and 497 nt of an open reading frame that extended the amino acid sequence amino-terminal to the tandem repeats by 165 residues before it ended at an exon-intron boundary. The open reading frame 5' to the repeats was different from that 3' to the repeats. The nucleotide sequence determined in these studies encoded the amino acid sequence from residues 1406 to 1570 in the mucin polypeptide.

Length of the Tandem Repeat Domain—With knowledge of the nucleotide sequence both 3' and 5' to the tandem repeat domains, it was possible to isolate clones from the genomic DNA library that hybridized with oligonucleotides containing sequences both 3' and 5' to the tandem repeats. Such clones would contain the entire length of the tandem repeat domain, and the number of repeats could be determined.

Three cosmids, designated Apo50, Apo60, and Apo70, were isolated that hybridized with oligonucleotides both 3' and 5' to the repeats. The inserts in these clones were removed by digestion of the clones with NotI, and agarose gel electrophoresis showed the lengths of the NotI fragments from Apo50, Apo60, and Apo70 to be 43, 45, and 34 kb, respectively, as shown in Fig. 2. Restriction maps of these cosmids for EcoRI and PstI are also shown in Fig. 3. Each clone has two EcoRI sites at the same position on either side of the DNA encoding the tandem repeats and one PstI site in each of the tandem repeats. Thus, as shown in Fig. 3, PstI digestion of each EcoRI fragment derived from a NotI fragment gives on gel electrophoresis a band of 243 nt encoding the tandem repeats and two other bands that are 1.7 and 2.4 kb in length, respectively. Previous studies (12, 13) showed that each 243-nt repeat contains one PstI site, which gives rise to the 243-nt band. Purification and sequence analysis of the 1.7- and 2.4-kb fragments showed the smaller fragment encoded sequence immediately 5' to the tandem repeats and the larger encoded sequence 3' to the repeats (12). Thus, the differing sizes of Apo50, Apo60, and Apo70 are the result of different numbers of tandem repeats they encode. The repeat domain of Apo50 encodes 125 repeats, that of Apo60 encodes 135 repeats, and that of Apo70 encodes 99 repeats, and each repeat domain is flanked by the 1.7- and 2.4-kb bands. These results show that porcine submaxillary mucin genes exhibit length polymorphism as a result of the different numbers of repeats they encode.

Fig. 3 also shows that Apo50 and Apo70 have no restriction sites for BamHI, although Apo60 had four such sites. This result shows that the tandem repeat domain displays site specific polymorphism as well as length polymorphism.

The above results also show that the genes encoding mucin contain no intervening sequences between the individual repeats, since PstI digestion produced only the tandem repeats and the unique sequences either 3' or 5' to the tandem repeat domain. Thus, an unusually long exon (34 kb to 32.8 kb) encodes the tandem repeat domain. The nucleotide sequence of the 135 tandem repeats in Apo60 encode the amino acid sequence of the mucin polypeptide from residues 1571 to 12529.

Determination of the Complete Sequence of the DNA 5' to the Tandem Repeats by Anchored PCR—The nucleotide sequence encoding the most amino-terminal portion of the mucin polypeptide was determined by anchored PCR with submaxillary gland RNA as template. The first cycle employed reverse transcriptase and primers with a sequence (bases 4674–4702) complementary to the unique sequence in Apo1 that was described above. A poly(G) sequence was added to the end of the single-stranded product from the first step and then amplified.

\(^2\) The complete nucleotide sequence and the derived amino acid sequence of porcine submaxillary gland mucin have been deposited in the GenBank™ data base under GenBank™ accession number AF005273. The electronic version of the Journal (URL: http://www.jbc.org) employs direct hyperlink access to entries in the data base, and this sequence can be generated by clicking on the above accession number. The sequences of human intestinal mucin (MUC2, GenBank™ accession number L21998) and human prepro-von Willebrand factor (GenBank™ accession number X04385), which have sequence identities with porcine submaxillary mucin, can also be generated by hyperlink.
in a thermocycler with an oligonucleotide primer complementary to the poly(G) sequence and another primer complementary to bases 4342–4362 that encodes the unique sequence 5’ to the tandem repeats. The amplified products were ligated into the pCRII vector and the cells screened with an oligonucleotide complementary with bases 4316 and 4335 of the DNA encoding the mucin as described under “Experimental Procedures.” The inserts in positive colonies were then sequenced. From the newly determined sequence, new primers were synthesized for additional cycles of anchored PCR. In all, a total of six cycles of anchored PCR were performed generating a continuous open reading frame of 4600 bp, which encoded the entire mucin sequence to the amino terminus, including the start site methionine. Immediately 5’ to the codon for the start methionine, stop codons were found in the DNA sequence in all reading frames. The methionine at the start site is in the sequence MKLIFYFIVALCFFC, and the 14 hydrophobic residues flanked by lysine residues, are thought to be the signal peptide for the secreted mucin. The site for proteolytic cleavage of the signal peptide has not been identified. The nucleotide sequence determined with aid of anchored PCR encodes the amino acid sequence of the mucin polypeptide chain from residues 1 to 1405.

The results of these studies showed that the porcine submaxillary mucin polypeptide chain had three disulfide-rich domains at its amino terminus that had considerable sequence identity with the D-domains of human intestinal mucin (Fig. 1) and human prepro-von Willebrand factor (21). Table I compares the amino acid sequence identity of the three D-domains of submaxillary mucin with the corresponding D-domains in human intestinal mucin and human prepro-von Willebrand factor. The table also lists the amino acid residues that are thought to comprise each of the D-domains of submaxillary mucin. Noteworthy is the absence of the D4-domain of human intestinal mucin in submaxillary mucin. The disulfide-rich domain at the carboxyl terminus of submaxillary mucin shows no statistically significant sequence identity with the D-domains, but has 30% identity with a similar domain at the carboxyl terminus of human intestinal mucin and 27% identity with the same domain in prepro-von Willebrand factor. However, 63% and 76%, respectively, of the half-cystine residues in this domain are in the same positions in the corresponding carboxyl-terminal domains of human intestinal mucin and human prepro-von Willebrand factor.

Confirmation of the 5’-untranslated sequence deduced as described above was obtained by anchored PCR with primers complementary to the nucleotide sequence at the extreme 5’ end of the gene as described under “Experimental Procedures.” Four of the eight inserts examined had sequences identical to nucleotides 1–94, three had the same sequence but started at nucleotide 2, and one had the sequence from nucleotide 24–94. These results showed that the 5’-untranslated sequence contains 55 nucleotides upstream from the initiation methionine.

The 5’-untranslated region has sequences for initiation of translation that are identical to the reported consensus sequence (22), except for an A substituted for a G at the +4 position.

Additional evidence that the full-length sequence of mucin
was obtained came from in vitro translation studies, in which a plasmid construct encoding the first 1360 amino acids of mucin was tested to determine whether the first methionine (at nucleotide 56 in the RNA sequence) could initiate translation after in vitro transcription. If initiation starts at the methionine, the plasmid should express a protein with Mr = 150,964. Gel electrophoresis of the expressed proteins showed only one species with an apparent Mr = 157,000 (data not shown), which is within 4% of the expected size. No proteins were observed when plasmid was omitted from the expression system.

Restriction Fragment Length Polymorphism of the Tandem Repeat Domain—The variability in the sequence of the tandem repeat domain in individual animals was evident from the Southern blots shown in Fig. 4. Fig. 4A shows the BamHI restriction patterns obtained with lymphocyte DNA isolated from four different pigs selected at random at the abattoir and employing the radiolabeled 243-nt tandem repeat sequence as the probe. Each differs from one another, although all have some bands that are identical to bands in another animal, suggesting considerable site-specific polymorphism in this domain of the mucin gene. Fig. 4B shows similar site-specific polymorphisms in the DNA from two families of pigs. In one family, only one of the progeny had a restriction pattern identical to one of the parents, and only two of the progeny had patterns identical to one another. In the second family, the restriction pattern of each family member differed from one another.

Tissue Distribution of Submaxillary Mucin RNA—Total RNA prepared from porcine tissues was screened in Northern blots with the radiolabeled 243-nt repeat as probe, as shown in Fig. 5. The autoradiograph was overexposed to detect small amounts of RNA. As expected, the submaxillary gland RNA hybridized most strongly to give a broad band with sizes ranging from 5 to 20 kb, as reported previously (13). The RNA from parotid glands and trachea also hybridized with the probe but not as strongly as the submaxillary RNA. A very small amount of RNA in large intestine also appeared to hybridize, but its nature in unknown. No hybridization was observed with the RNA from other tissues that secrete mucins including sublingual gland, oral cavity mucosa, stomach, small intestine, uterus, and pancreas. It is also not known whether the trachea and parotid glands make submaxillary mucin RNA, or RNA for a different mucin structurally similar to the submaxillary RNA. Nevertheless, the submaxillary gland mucin seems to have a limited distribution among porcine tissues.

DISCUSSION

Because of the great length of the DNA encoding the mucin polypeptide chain, it was necessary to use three different techniques to obtain the full-length sequence. Expression cloning of a agt11 library screened with anti-mucin polypeptide chain antibodies gave cDNA that encoded the carboxyl-terminal disulfide-rich domain, about five tandem repeats, and the sequence that joins the disulfide-rich domain and the tandem repeats (12, 13). Screening of a porcine genomic cosmid library was required to obtain DNA sequences that contained the entire repeat domain and the sequence immediately 5' to the repeat domain. With this information, the remaining sequence from the tandem repeat domain to the initiation site was obtained. Fortunately, the tandem repeat domain is encoded by DNA devoid of intervening sequences; thus, the number of repeats in the domain was more easily determined. Preliminary studies suggest that many introns occur in the regions coding the other domains of mucin, but the complete structure of the entire genomic DNA encoding porcine submaxillary mucin has not been determined.

Based on previous studies (12, 13), the tandem repeat domains of the mucin polypeptide chain and the sequences that join this domain to the domains at either end of the molecule are thought to contain the O-linked oligosaccharides that comprise about two-thirds of the weight of the native mucin. This has been firmly established by recent studies (23), which demonstrate that each of the 19 serine and 11 threonine residues in this domain contain O-linked oligosaccharides. The largest form of the polypeptide chain of mucin has an Mr = 1,184,106 and would be about 3,500,000 if fully glycosylated. Light scattering studies (24) indicated that submaxillary mucin had a Mr = 2-3 million in 6 M guanidine hydrochloride and decreased to about 900,000 in the presence of mercaptoethanol. It is now known that the mucin chains likely form dimers through in-
terchain disulfide bonds in the carboxyl-terminal disulfide-rich domain and possibly higher oligomers through disulfide bonded amino-terminal D-domains (11). Thus, the observed values for the molecular weights are somewhat smaller than expected from the studies described here. The discrepancies could be explained if the mucin used for light scattering was proteolytically degraded, as can easily happen (13, 25), or contained a tandem repeat domain somewhat shorter than those found here.

As noted previously, the tandem repeat domain and its flanking sequences that are similar in composition to the tandem repeat domain appear to be devoid of secondary structures (13, 25). Thus, this domain is thought to serve solely as a scaffold for the O-linked oligosaccharides (25), but because of charge repulsion among the negatively charged sialic acids in the oligosaccharides, it forms a semi-rigid, extended structure of the kind observed by light scattering (26) and electron microscopy (27). There are no NX(S/T) sequences in the tandem repeat, but there are a total of 18 in the mucin polypeptide. Four of the asparagines (residues 13056, 13123, 13140, and 13206) are likely glycosylated in the carboxyl-terminal disulfide-rich domain (11), but it is not known whether those in the D1, D2, and D3 domains (residues 175, 260, 524, 547, 741, 825, 916, 961, 1292, 1364, and 1463) are glycosylated. There are also three asparagines at residues 12556, 12685, and 12953 that are in the sequence NX(S/T). However, there are a total of 4050 serine and threonine residues in the tandem repeat domains alone, so the contributions of N-linked oligosaccharides to the viscosity of mucin are likely very small.

Three different forms of the gene encoding mucin were isolated from a cosmid genomic DNA library. The library was prepared from the DNA from a single animal, and not pooled DNA. Each form was identical except for the number (99, 125, and 135) of repeats in the tandem repeat domain. An exhaustive search for additional genes with different numbers of repeats was not made, but our finding of only three different genes is consistent with previous studies (28) on the genetic analysis of porcine submaxillary mucin showing that it is controlled by a three-allele system with codominant inheritance. Rearrangement and/or recombination of DNA encoding human intestinal mucin have been reported in bacteria (29); however, there is no evidence that this was the case here. Clones with DNA inserts of porcine mucin were stable over more than 3 years and showed no structural changes during cloning and manipulation. Possibly, there is some somatic polymorphism within an individual at a given locus, but this cannot be determined from the family data (Fig. 4). It is not known whether the expression of the different forms is tissue-specific, although this remains to be established. The formation of disulfide bonds of von Willebrand factor is dependent on the tetrapeptide motif CGLC (32), in the D-domains. Interchain disulfide bonding of von Willebrand factor is dependent on the tetrapeptide motif CGLC (32), in the D-domains. Interchain disulfide bonds of von Willebrand factor forms disulfide-bonded multimers through interchain disulfide bonds in its amino-terumoinal D-domains (10), suggesting that the analogous domains in mucin possibly aid assembly of mucin dimers into multimers, although this remains to be established. The formation of disulfide bonded oligomers of von Willebrand factor is dependent on the tetrapeptide motif CGLC (32), in the D-domains. Interestingly, the same motif occurs twice in the D-domains of submaxillary mucin, in about the same positions as in von Willebrand factor. If oligomers of submaxillary mucin, or other secreted mucins, are formed in much the same manner as in von Willebrand factor, then their exact cytoprotective functions remain to be established. However, these structure-function relationships between von Willebrand factor and the secreted mucins argue strongly that the D-domains in secreted mucins and von Willebrand factor are homologous structures and are evolutionarily related through a common ancestral gene (1). Similarly, the genes encoding the disulfide-rich carboxyl-terminal domains in these proteins appear to be derived from a common ancestral gene.
